ESD-TDR-64-601

(FINAL REPORT)

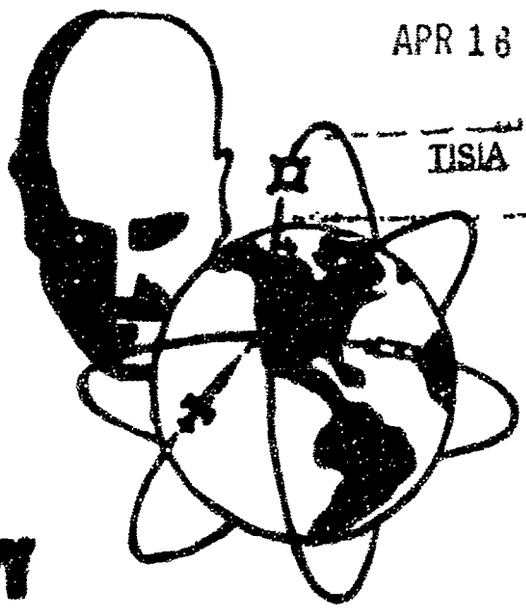
HUMAN PROCESSING OF EQUIVOCAL INFORMATION

TECHNICAL DOCUMENTARY REPORT NO. ESD-TDR-64-601

April 1965

Ward Edwards

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts



Project 4690, Task 469003

(Prepared under Contract No. AF 19(604)-7393 by the Engineering Psychology Laboratory, Institute of Science and Technology, The University of Michigan, Ann Arbor, Michigan.)

DDC AVAILABILITY NOTICES

3780-23-F

HUMAN PROCESSING OF EQUIVOCAL INFORMATION

Final Report

Ward Edwards

April 1965

Engineering Psychology Laboratory

INSTITUTE OF SCIENCE AND TECHNOLOGY
The University of Michigan
Ann Arbor, Michigan

FOREWORD

This final report covers three years of work in the Engineering Psychology Laboratory at the Institute of Science and Technology of The University of Michigan. The work was performed for the Operational Applications Laboratory of the Electronics Systems Division of the Air Force Systems Command, and was conducted in accord with the terms of United States Air Force Contract AF 19(604)-7393. Contracts and Grants to The University of Michigan for the support of sponsored research by the Institute of Science and Technology are administered through the Office of the Vice-President for Research.

The author wishes to acknowledge the contributory work of M. Guyer, W. L. Hays, R. Norman, L. D. Phillips, S. M. Rubin, M. A. Swain, and M. T. Zivian; and to indicate his gratitude to J. T. Begley, Col. A. Debons, G. P. Mandanis, W. E. Organist, L. J. Savage, E. H. Shuford, Jr., A. W. Story, and D. H. Wilson for information, advice, and criticism.

This is Institute of Science and Technology Report Number 3780-23-F.

REVIEW AND APPROVAL

This Technical Documentary Report has been reviewed and is approved

DONALD W. CONNOLLY
Chief, Display Division
Decision Sciences Laboratory

ROY MORGAN, Col, USAF
Director
Decision Sciences Laboratory

CONTENTS

FIGURES

TABLES

HUMAN PROCESSING OF EQUIVOCAL INFORMATION

ABSTRACT

This report contains a series of studies investigating the abilities of subjects to revise probability estimates on the basis of new information. These studies show that subjects' probability estimates are reliable but deviate considerably from posterior probabilities calculated from Bayes's theorem. The deviations are almost always in the conservative direction, i.e., low Bayesian probabilities are overestimated, and high ones are underestimated. Only when each datum is very ambiguous do subjects' estimates become more extreme than Bayesian probabilities. Further, when subjects are asked to give 90% or 50% credible intervals of a posterior probability distribution, their estimates are wider than Bayesian credible intervals. This finding of conservatism has led to the design of a man-computer system that should minimize the effects of human shortcomings in making diagnoses.

1 INTRODUCTION: A SCIENTIFIC OVERVIEW[1]

This is the final report of a three-year program of research into human information processing and decision making, applying techniques based on Bayes's probability theorem to the design of man-machine systems for information processing (Bayes's theorem is explained in detail in Section 2). Appendix B briefly summarizes the publications already in print or in press that have developed from the contract. In order to keep this final report to reasonable length, no attempt will be made to duplicate information contained in publications summarized in Appendix B; the primary purpose here is to report research completed under Contract AF 19(604)-7393 but not yet published.

Certain activities conducted under this contract cannot be properly reflected in a final report. One of them is the development of research plans for semisimulation experiments concerned with probabilistic information-processing systems. These research plans occupied a great deal of time and attention during the last 18 months of the contract, but have by no means reached fruition as yet. This work is being continued under Contract AF 19(628)-2823, and will appear in publications sponsored by that contract.

A second class of activity that cannot be adequately reflected in this final report is the proceedings of a conference on Bayesian Information Processing Systems held at The University of Michigan in May, 1963. Participants in this conference exchanged information about research on men as Bayesian information processors and about the design and evaluation of Bayesian

[1] This Section was prepared by Ward Edwards.

information-processing systems. No formal publications were intended to result from this conference; its purpose was to facilitate informal interchange of information and to bring different researchers concerned with related problems into interaction with one another so that their research would coordinate and make a more cohesive whole than might otherwise be possible.

Research conducted under Contract AF 19(604)-7393 was of three major kinds. One consisted of primarily theoretical investigations into the formal characteristics of Bayes's theorem as a mathematical model for the revision of opinion in the light of information. This work culminated in a long article about the relevance of Bayesian ideas to statistics, another concerning the optional stopping problem, and several minor efforts. A second kind of research consisted of laboratory studies comparing man and Bayes's theorem as information processors, and finding that man is the more conservative; a number of studies elaborated this finding and examined some of the parameters that affected it. The third kind of endeavor under the program was the development, elaboration, and thinking-through of an idea for a Bayesian information-processing system, or PIP, followed by the development of semisimulation research techniques for the exploration and validation of that idea. Of the three classes or research, this one had the least visible product, since it consisted primarily of intellectual effort, mostly of a nonpublishable nature. Nevertheless, this class of endeavor seems likely to have the greatest impact in the long run on military technology and Air Force system design.

This scientific overview, which is really nothing more than a brief introduction both to the publications that have already emerged from this contract and to the chapters that follow, will ignore the first kind of effort completely. The formal work on Bayesian statistics and optional stopping has been fully reported in publications, stands on its own feet, and needs neither amplification nor review. The summaries of publications in Appendix B briefly report what was done.

The main research endeavor of the contract was concerned with the comparison of men with Bayes's theorem as probabilistic information processors. When the contract began, essentially no information about the quality of human information processing in unspeeded tasks was available. It was widely supposed that men were good information processors, but little was known about how good, mostly because there was no formal model for proper information-processing methods. The research started from the premise that Bayes's theorem was an optimal model for information processing, and consequently that straightforward experiments comparing human performance with the output of Bayes's theorem might lead to insights regarding the quality of human information processing. Thus the experiment described in Section 2 was designed as a frontal attack on the problem. It used a very complicated task involving 4 mutually exclusive hypotheses and 12 different possible observations, displayed to the subjects a set of 48 probabilities of the data given the hypotheses, and then required the subjects to generate posterior probability estimates. Not too surprisingly, it turned out that their estimates differed from

Bayesian probabilities. What was more interesting, however, was that these differences were consistently in the direction that we have come to call conservative; that is, subjects consistently overestimated low Bayesian posterior probabilities and underestimated high posterior probabilities. No subject extracted from the data anything approaching the certainty it would justify.

It seemed entirely possible to us that these deviations from Bayesian probabilities, overwhelming in size and consistency though they were, might have been attributable to artifacts of one kind or another. Consequently, we designed two experiments intended to examine two artifacts that we thought might be relevant—and happily found that neither artifact need be considered too seriously. In one of these experiments (Section 3) we asked whether subjects might have been confused by the fact that in the original experiment the data did not resemble any of the distributions of data to be expected with the given hypotheses; it turned out that this made no difference whatever. In the other experiment we examined the effects of sequential and nonsequential presentation of data and found much the same amount of conservatism, whether the subject in effect started from scratch each time or was allowed to retain his previous posterior probability estimates for use as prior probabilities, with only an incremental datum added.

In another study, (H. C. A. Dale, unpublished), subjects were allowed to specify their own values of the probability of the datum, given the hypothesis [P(D|H)]; still they were conservative. It seems that whether or not the value of P(D|H) conforms to the subject's intuitive appraisal of what it ought to be makes very little difference to his information-processing performance.

Still another study (Section 3) raised the question of whether it is ever possible to get subjects to overestimate a posterior probability. It turns out that the answer is yes, if the information given to the subject is sufficiently worthless for diagnostic purposes. That is, when Bayesian posterior probabilities are very little different from Bayesian prior probabilities, a subject's estimates of posterior probabilities usually are more extreme.

Next we turned our attention to the question of whether this conservatism could be found also in much simpler, more straightforward kinds of experiments. In one such experiment (Section 4), we presented subjects with observations drawn from a normal distribution and required them to estimate a posterior credible interval for the mean. Findings from this study were entirely consistent with the findings from the previous, more complicated study: subjects were consistently conservative. The task of estimating a credible interval, however, is unfamiliar to subjects and their estimates were rather variable.

Finally, we sought the simplest possible task in which subjects could perform this kind of information processing (Section 5). We ended by using a simple binomial task in which subjects must decide which of two hypotheses about the percentage of red poker chips in a bookbag full of red and blue poker chips is correct. Here, too, we obtained conservatism, though not quite so much of it as in the experiment reported in Section 2. The data indicated very clearly that

even in this simplest of all possible Bayesian tasks, subjects are unable to extract from information all the certainty that is latent in it. This situation seemed appropriate for further study, so we designed a number of experiments, many of them still incomplete, examining various of its parameters. One, sufficiently complete for inclusion in this report (Section 6), compared three different modes: estimating probabilities on a device displaying a linear scale of probabilities, estimating odds verbally, and estimating odds on a device displaying a logarithmic scale of odds. It had seemed possible that one reason for conservative performance was that the probability scale is bounded at zero and one, and subjects are consequently reluctant to come too close to the boundaries at which they have no more freedom to move. However, the experiment on response modes indicates clearly that this factor, while relevant, is not the primary cause of conservatism. The two odds groups show less conservatism than the probability group but still plenty of it; the logarithmic scale seems to produce very slightly less conservatism than the direct verbal reporting of odds. Research of this kind continues.

The fundamental finding of the first study has required no qualification or modification as a result of its amplification by these further experiments. The basic bias seems to be strong and very nearly universal (at least among college students), although of course the magnitude of the effect is influenced by a variety of such peripheral factors as response modes, presence or absence of payoffs, complexity of the task, amount of training received, presence or absence of feedback concerning the correct hypothesis, etc. And it is appropriate to ask what effects this consistent bias in human behavior might have on such practical problems as the design of information processing systems.

Existing systems intended for processing information in decision making, such as command and control systems, may be extremely sophisticated in their information gathering, display, and communications. But their technique for processing the information obtained is identical with that used by Alexander the Great: display it to the commander and let him decide. Clearly, any bias that the commander may bring to his process of deciding will be a bias in the operation of the system. It seems very likely, in view of the research findings and also on intuitive grounds, that commanders have a conservative bias in such systems; that is, that they are unable to extract all the certainty from the data that the data would justify. Therefore one problem in the design of information-processing and decision-making systems may be defined as the problem of how to prevent the natural conservatism of human information processing from making such systems less responsive than they should be.

One step toward the solution of that problem consists of analyzing information-processing into subtasks. It seems clear that at least two such subtasks can be discriminated. One consists of assessing the impact of a single item of information on some hypothesis, or set of hypotheses, of interest to the system. The other consists of aggregating these impacts over data and over hypotheses into a picture of the current status of the hypotheses. The first of these tasks, for the kinds of qualitative information that are characteristically available to information-processing

systems, must inevitably be performed by expert human beings. But the second of these tasks is naturally performed by Bayes's theorem and consequently is easy to mechanize.

The essence of the proposal for the design of a probabilistic information-processing system that has emerged from the research of this project is that human beings should estimate the probability of each datum given each hypothesis — P(D|H) — for some closely related quantity, such as a set of likelihood ratios, and that a machine should be used to aggregate these estimates into a posterior distribution over the hypotheses of interest to the system.

If human estimators are naturally less conservative in estimating likelihood ratios than in estimating posterior probabilities, or can be taught to be so, then this procedure should cope with the problem of conservative bias. In any case, it seems attractive on other grounds: it permits fragmentation of the task of information-processing into many subtasks that can be parceled out among men and machines in a manner respecting the capabilities of each. Moreover, it permits full mechanization of what is, from the human point of view, basically a bookkeeping task: the aggregation of data into posterior distributions.

The research problems of specifying and evaluating this idea are numerous and very difficult. One problem asks how men can be selected, trained, and provided with suitable displays and controls so that they can work effectively as estimators for P(D|H) or a related quantity. Another, which assumes that trained men can provide suitable estimates, asks how a system can be designed to exploit that fact. Only the first of these two questions has been of primary interest in the research program of this contract; the other is the business of Contract AF 19(628)-2823.

Simple, short-term, inexpensive laboratory experiments are incapable of studying probability estimation in really complex situations under full experimental control. Either the expertise of the subjects and the context in which they estimate probabilities must be artificially created in the laboratory, in which case the expertise cannot be very deep nor the context complicated, or else contexts and abilities pre-existing in the real world must be studied. The former procedure obviously falls short of examining performance under realistically complex circumstances. The latter procedure sounds more attractive. Unfortunately, it is almost impossible to determine the "correct" probabilities in real-world contexts. It is also nearly impossible to insure that different subjects have comparable amounts of information about the real-world contexts chosen for study. So the results obtained from the use of real-world contexts and abilities would be hard to interpret, especially if the question asked is how "correct" these estimates are.

One product of the research program of the contract is a proposed solution to these difficulties. The proposed solution is expensive, time-consuming, and difficult, but it may work. It is to synthesize a partially artificial world, of the greatest complexity consistent with tractability, that has well-specified probabilities built into it. Once this complex artificial world has

been invented, it is necessary to train subjects to be expert about it—a long process. After that has been done, they can be exposed to information-processing tasks appropriate to that artificial world, and asked to estimate P(D|H) or similar quantities for suitable data and hypotheses. Both their estimates and the performance of the probabilistic information-processing system that uses them can be evaluated by comparison with the "true" probabilities built into the world to start with, and perhaps also by comparison with the performance of nonprobabilistic systems in the same setting.

The semisimulation research planned under this contract will do just that.

2
CONSERVATISM IN COMPLEX PROBABILITY INFERENCE[2]

This experiment examines the relationship between estimated probabilities and probabilities calculated by means of Bayes's theorem; it compares human performance with optimal performance in the task of revising opinion in the light of new information.

To provide setting and vocabulary, a number of contemporary ideas concerning probability must be briefly summarized. The numbers called probabilities are formally well-defined by the assertions that they are numbers between zero and one, and that over a mutually exclusive and exhaustive set of hypotheses (hereafter called a partition) they must add to one. But three fundamentally different operations have been proposed to relate those numbers to observable events in the real world. The currently dominant frequentistic view defines a probability as the limit of the relative frequency with which a particular phenomenon occurs; probabilities can be estimated, for example, by operations like tossing a coin many times under "substantially equivalent" conditions, and then using the ratio of heads to total tosses as an estimate of the probability of heads. The symmetristic view appeals to observable symmetries to make plausible the notion of a collection of equally likely elementary events; the faces of a die are considered equally likely to come up because the die is symmetrical. The personalistic view defines a probability as an orderly and coherent judgment made by a rational person who brings to bear upon the immediate question his relevant past experience, of whatever kind. Probabilities so defined are called personal probabilities, and describe the person judging the event as well as the event itself.

Corresponding to these three philosophical positions about the foundations of probability are three quite different ways of displaying probabilities. Symmetry displays are common and effective; examples are cards, dice, roulette wheels, and the like. Frequency displays are very rare, mostly because frequencies are usually based on counts of random samples, and the notion of randomness is usually defined by an appeal to symmetry. But what might be called plausibility displays are most common of all. We make intuitive nonnumerical judgments of probability at every moment of our lives, and any information display that influences such judgments without necessarily appealing to symmetry or relative frequency (or mathematical necessity) may be called a plausibility display.

Philip [1], Stevens and Galanter [2], and Shuford [3] have found that simultaneously displayed relative frequencies can be quite accurately estimated on the basis of exposures too short to permit counting, and Robinson [4] found the same thing for sequentially displayed relative frequencies. Teichner [5], in a more complex task using a frequency display, obtained somewhat less accurate performance.

[2]This section was prepared by Lawrence D. Phillips, William L. Hays, and Ward Edwards.

A large number of experiments have attempted to infer judged probabilities from observed acceptances and rejections of bets, assuming that subjects based their decisions on some version of the well-known Subjective Expected Utility (SEU) model. (For reviews of this literature and of the model, see Edwards [6, 7].) Such studies typically use symmetry displays of probability such as are provided by dice, spinners, and the like. In effect, then, probabilities inferred from decisions via the SEU model are compared with probabilities displayed directly by means of symmetry displays. The large systematic differences that are almost always found in such studies imply either that symmetry displays produce severe distortions of judged probabilities (compared with the generally accurate judgments of relative frequencies, for example) or else, more plausibly, that the SEU model is descriptively inadequate and so is an inappropriate basis for inference of judged probabilities. But methodological and formal difficulties dominate this literature and few firm conclusions are possible.

Some of the descriptive inadequacies of the SEU model can be alleviated by using a conception of probability that does not require the sum of the probabilities of a mutually exclusive and exhaustive set of events to be one, or any other constant. Whether such numbers deserve to be called probabilities could be argued, but they can be so considered, and a nonadditive SEU model is not internally contradictory (Edwards [8]), although in such a model, utilities must be measured on a ratio, rather than an interval, scale. Such possibly nonadditive probabilities inferred from the choices of real people might well be called subjective probabilities, to distinguish them from the personal probabilities that might be inferred from the choices of ideally consistent people.

The clouds on Venus either contain a lot of water vapor or they do not. For a frequentist, the probability of that proposition is therefore either one or zero, if it is defined at all. A personalist, however, prefers to express his uncertainty about the clouds on Venus (and indeed about any topic) as explicitly as possible, and uses probabilities to do so. He consequently considers the probability that the hypothesis is true—a notion meaningless to frequentists.

Bayes's theorem, an elementary and noncontroversial consequence of the definition of conditional probability and of the requirement that probabilities must add up to one over a mutually exclusive and exhaustive set of events, has some usefulness for frequentists. For personalists, however, it plays a crucial role: it is the formally appropriate rule specifying how the probability that a hypothesis is true should be revised in accord with new data. It is therefore an optimal model for revision of opinion in the light of information—that is, for information processing.

Bayes's theorem can be expressed as follows:

$$P(H|D) = kP(D|H)P(H) \qquad (1)$$

$P(H|D)$ is the probability assigned to hypothesis H, given knowledge of the datum D; $P(H)$ is the probability assigned to H before D was known; and $P(D|H)$ is the probability of getting data if H is true. The normalizing constant, k, ensures that

$$\sum_{i=1}^{m} P(H_i|D) = 1 \text{ over the m elements of the partition}$$

It is easy to show that

$$1/k = P(D) = \sum_{i=1}^{m} P(D|H_i)P(H_i)$$

$P(H|D)$ is called the posterior probability of H, and $p(H)$ is the prior probability: $P(D|H)$ is called the likelihood (of datum D on hypothesis H). Under circumstances such as those prevailing in our experiment, the likelihood of several data is simply the product of the individual likelihoods. Formally, this simple rule is appropriate only if the data are conditionally independent of one another given each of the hypotheses; for a discussion of the difficult topic of conditional independence, see Edwards, Lindman, and Savage [9].

Thus, Bayes's theorem says that the probability assigned to a hypothesis after observing the datum (or data) D is directly proportional to the probability assigned to the hypothesis before observing the datum multiplied by the likelihood of the datum.

These experiments compared the posterior probability estimates of several subjects with the probabilities calculated by means of Bayes's theorem and investigated several variables that affect posterior estimates. Subjects were told that the artificial environment for this experiment could be in exactly one of four states, referred to as hypotheses, and that they would observe data generated by only one of these hypotheses. Subjects were shown the values of the individual likelihoods, that is, $P(D|H)$ for each possible datum, and were given the prior probabilities assigned to the hypotheses before observing any data. Then, subjects were asked to revise their opinions about which hypothesis was true after each new datum. However, subjects were not allowed to make any computations. They were required simply to make intuitive estimates of the posterior probabilities. Since the only constraint placed on subjects was that their estimates be between zero and one, the posterior estimates can be considered subjective probabilities. Personal probabilities were calculated from Bayes's theorem using the given prior probabilities and likelihoods.

2.1. EXPERIMENT ONE

2.1.1. METHOD

2.1.1.1. Procedure. Each subject was seated at a console and asked to imagine himself at the output of a large, computerized radar system. Subjects were told that the environment

in which this system operated was in one of four states: enemy attack, friendly activity, meteor shower, or enemy attempt to spoof the surveillance system. The system detected aerial activity and computed the predicted points of impact of the objects detected. These points, the data of the experiment, were displayed to the subject on a representation of a circular land mass that had been divided into twelve sectors. (A sample display is shown in Fig. 1.) Impact points always appeared within the sectors, never on the sector borders. These displays were projected from a 35-mm slide projector onto the rear of a rectangular viewing screen, 12 by 8 inches, located on the console slightly above eye level when the subject was seated.

After each display the subject estimated the posterior probabilities that the system was detecting each of the four kinds of activity. Their estimates were made by setting four levers mounted at five-inch intervals on the sloping front panel. Each lever had a 12-inch travel with the 0 setting nearest the subject, the 1.0 setting furthest from the subject, and calibration marks every 0.05.

To help him make these estimates, the subject was given the prior probability for the Enemy hypothesis and all possible values of $P(D|H)$. The prior probability was displayed above the unused response levers. The displays of $P(D|H)$ for each of the four hypotheses were located above the middle four response levers. This row of displays and the response levers are shown in Fig. 2 (the two outside levers were unused). The probabilities shown are the ones used in this experiment.

The subject was told that the display of $P(D|H)$ for a particular hypothesis represented the probabilities that the impact points would fall in the corresponding sectors if in fact that kind of activity was occurring. He received no instructions about how to use these numbers, except the obvious qualitative statements, and he was not told that the likelihood of several dots is equal to the product of the likelihoods for the individual dots.

After making his estimates for a display, the subject pressed a button that instructed the machine to record them; after that, he reset his levers to zero before seeing the next display. No constraint was placed on the sum of the posterior probability settings; subjects who asked were told the sum was up to them. Subjects gained familiarity with the apparatus during the instruction session and during the subsequent trial run. Subjects were never told anything about the quality of their estimates. (Complete instructions are in Appendix A.)

2.1.1.2. Stimuli. Each subject was presented with 32 ordered sequences of 15 stimulus slides each, and with 32 scrambled sequences constructed from the ordered sequences. The first slide in an ordered sequence contained only one dot (impact point), the second showed the

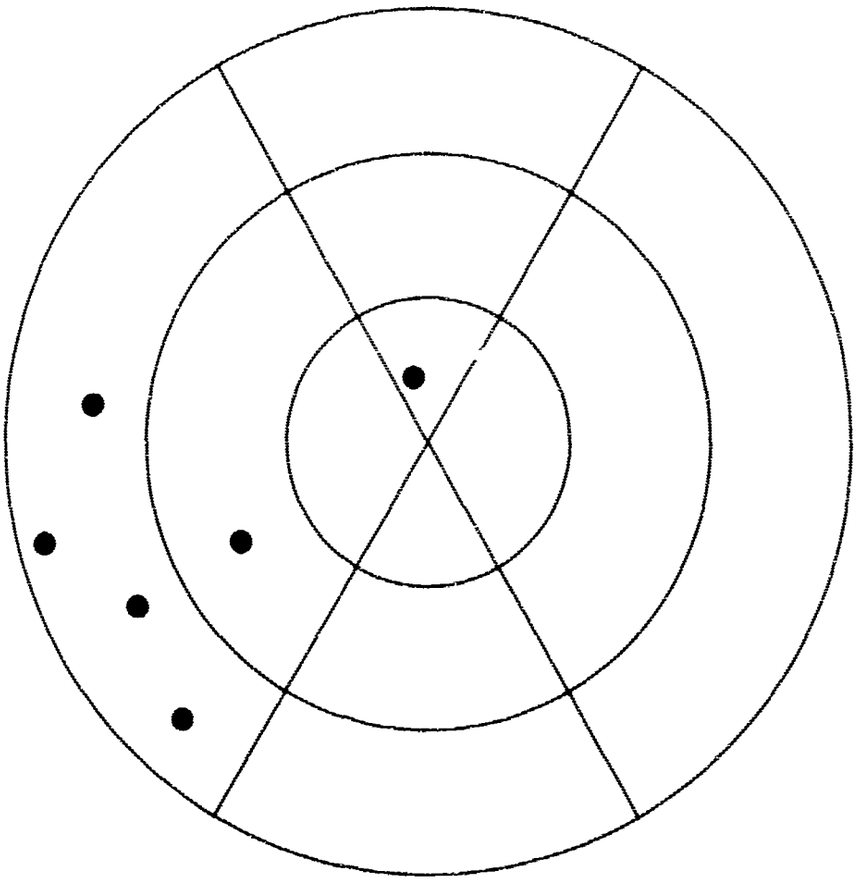


FIGURE 1. A SAMPLE DISPLAY OF SIX IMPACT POINTS

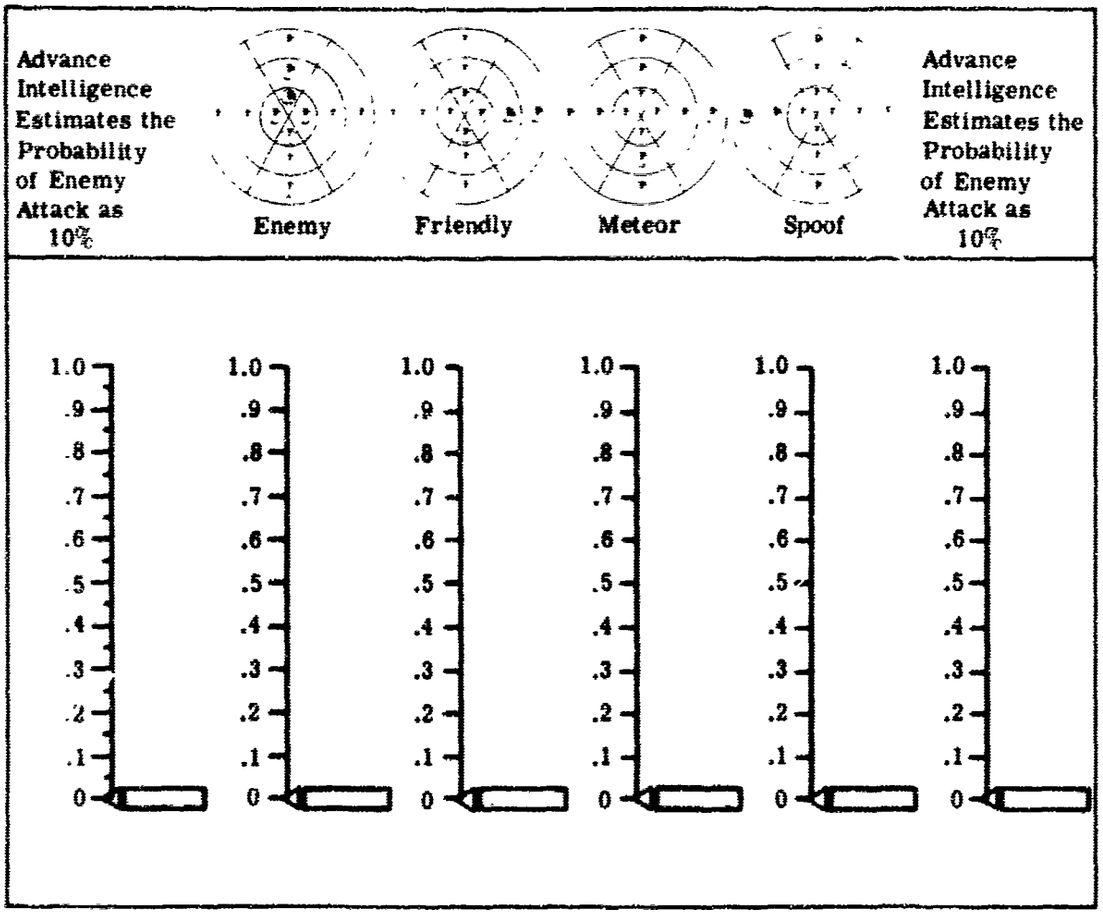


FIGURE 2. PRIOR PROBABILITY STATEMENTS, CONDITIONAL PROBABILITY DISPLAYS, AND RESPONSE LEVERS

first dot plus a new one, the third slide contained the first two dots plus a new one, and so forth for the remaining slides. Each ordered sequence was designated with a two-digit number.

The scrambled sequences were constructed by mixing two ordered sequences together and drawing at random without replacement two new sequences of fifteen slides from the total set of thirty slides. The scrambled sequences, then, showed no orderly accumulation or progression of dots. Each sequence, ordered or scrambled, had one of three Enemy prior probabilities associated with it — 10%, 25%, or 67%. Plots of the theoretical (Bayesian) posterior probabilities for three representative sequences are shown in Fig. 3. Bayesian probabilities were computed using the Enemy prior probability, which was given, and assuming that the remaining prior probability was distributed equally over the other three hypotheses.

For any given ordered sequence, the dots fell into exactly three of the 12 sectors, but the three sectors used were not necessarily the same from sequence to sequence.

To summarize, three variables were investigated: amount of information (number of dots) prior probabilities and order of presentation of information (ordered vs. scrambled).

Subjects participated in two-hour sessions, during which six to eight sequences were usually completed, until all 64 sequences had been shown. The total time a subject needed to complete all sequences varied from 14 to 25 hours. The order of presentation of the stimulus sequences was partially counterbalanced over the five subjects by use of a lattice design (Cochran and Cox [10]).

2.1.1.3. Subjects. Five volunteers, male University of Michigan freshman engineering students, were paid at the rate of $1.25 per hour. This population was chosen to insure familiarity with quantitative reasoning and ignorance of Bayes's theorem.

2.1.2. RESULTS. It is convenient to discuss several minor findings first, since they permit great simplification of analyses of the major finding.

2.1.2.1. Sum of Posterior Settings. One subject spontaneously attempted to constrain the sum of his posterior settings. Another asked if his settings should sum to one, but when told that he could do as he liked, did not normalize his settings. The remaining three subjects did not normalize their settings. For these latter four subjects, the sums of their posterior probability settings increased with the number of stimulus dots. None of the other variables had any consistent influence on this sum. Introspection and inquiry suggest that the main reason for this is that subjects are much more willing to increase an estimate for a diagnosis favored by a new item of evidence than to decrease estimates for the diagnoses not favored by that item. Plots of the sums for each subject are shown in Fig. 4. These sums are averages over all 64 stimulus sequences.

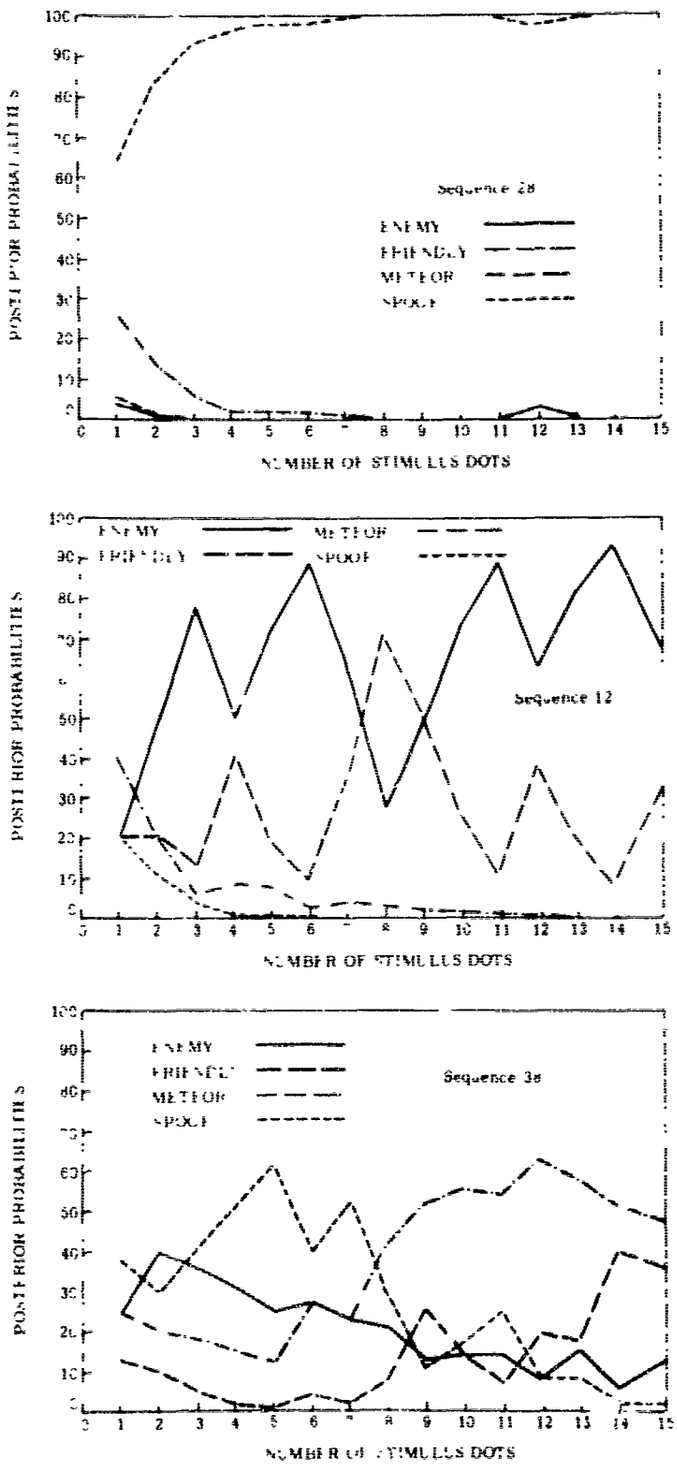


FIGURE 3. PLOTS OF BAYESIAN POSTERIOR PROBABILITIES

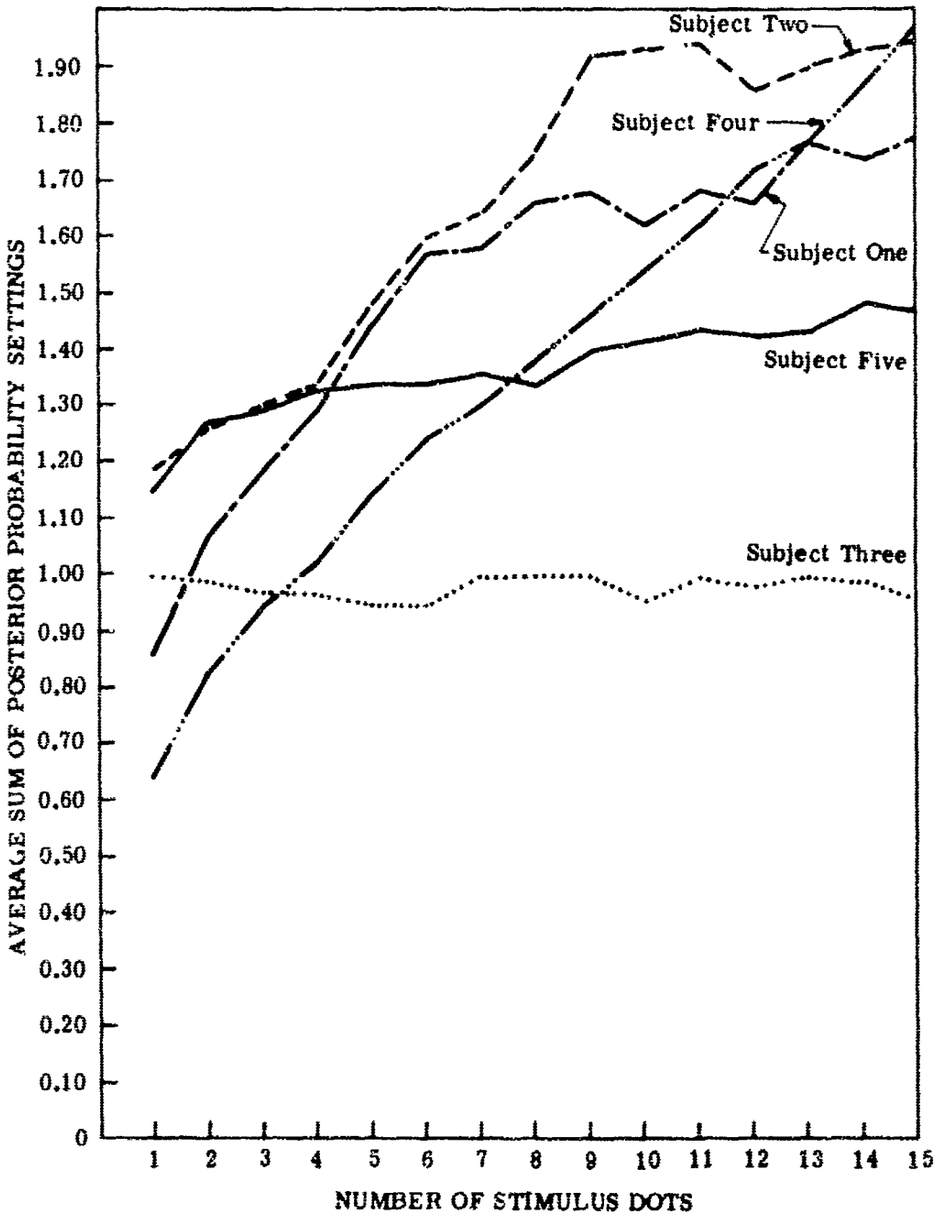


FIGURE 4. AVERAGE SUM OF POSTERIOR PROBABILITY SETTINGS AS A FUNCTION OF NUMBER OF STIMULUS DOTS

2.1.2.2. Analysis of Variance of Deviations. Squared deviations of subjects' posterior estimates of enemy attack from the theoretical Bayesian values were computed. The mean over the number of dots of the squared deviations was defined as a measure of the amount of deviation from optimal performance. Analysis of variance of this measure showed that the main effect of prior probabilities was significant beyond the 0.01 level. The order of presentation of information showed absolutely no effect; the interactions of order with prior probability and with individual sequences were insignificant. For that reason, subsequent data analyses combine data obtained from both ordered and scrambled conditions, or else consider only the ordered condition.

Lack of significance of the order variable is surprising. Apparently, subjects' deviations from optimality are unaffected by the order in which they receive information. In this respect, subjects' behavior is like that of Bayes's theorem. In order to compute posterior probabilities for any given slide, the theorem needs to know only the conditional probabilities of observing all the data displayed, and the prior probability that obtained before the data were observed. These probabilities can be obtained without knowledge of any other slides. We conclude, then, that for this task, subjects are little affected by the sequential nature of the information in the ordered sequences; each slide is treated as a separate problem.

To facilitate more meaningful analyses of the data, subjects' posterior estimates were adjusted proportionately so that the sum over the four hypotheses was one. Analyses in the remainder of this report use only the normalized data.

2.1.2.3. Deviations from Bayes's Theorem. Figure 5 shows representative plots of subjects' estimates (after normalization) as a function of the number of stimulus dots. These estimates should be compared with the Bayesian probabilities shown in Fig. 4. The lack of any systematic difference between ordered and scrambled presentation is evident. But the most striking finding is the very small amount that subjects changed their probability estimates from one stimulus to the next, even when Bayesian probabilities showed considerable change. In nearly every sequence, subjects exhibited this conservatism. Subject Three is the only exception: he sometimes moved more than Bayes's theorem. In some cases, notably for Subject Four, the posterior estimates moved toward one another instead of toward zero or one as the number of dots increased. This subject apparently became less sure as the evidence mounted up. Even on problems as easy as the top one in Fig. 4 and 5, four of the five subjects failed to reach anything like the extreme posterior probabilities that would be appropriate. Subject Three, the most nearly Bayesian subject throughout the experiment, did better than any of the others, but still not well. Though the details vary from sequence to sequence and from subject to subject, the finding is the same for nearly all: subjects failed to be as sure as Bayes's

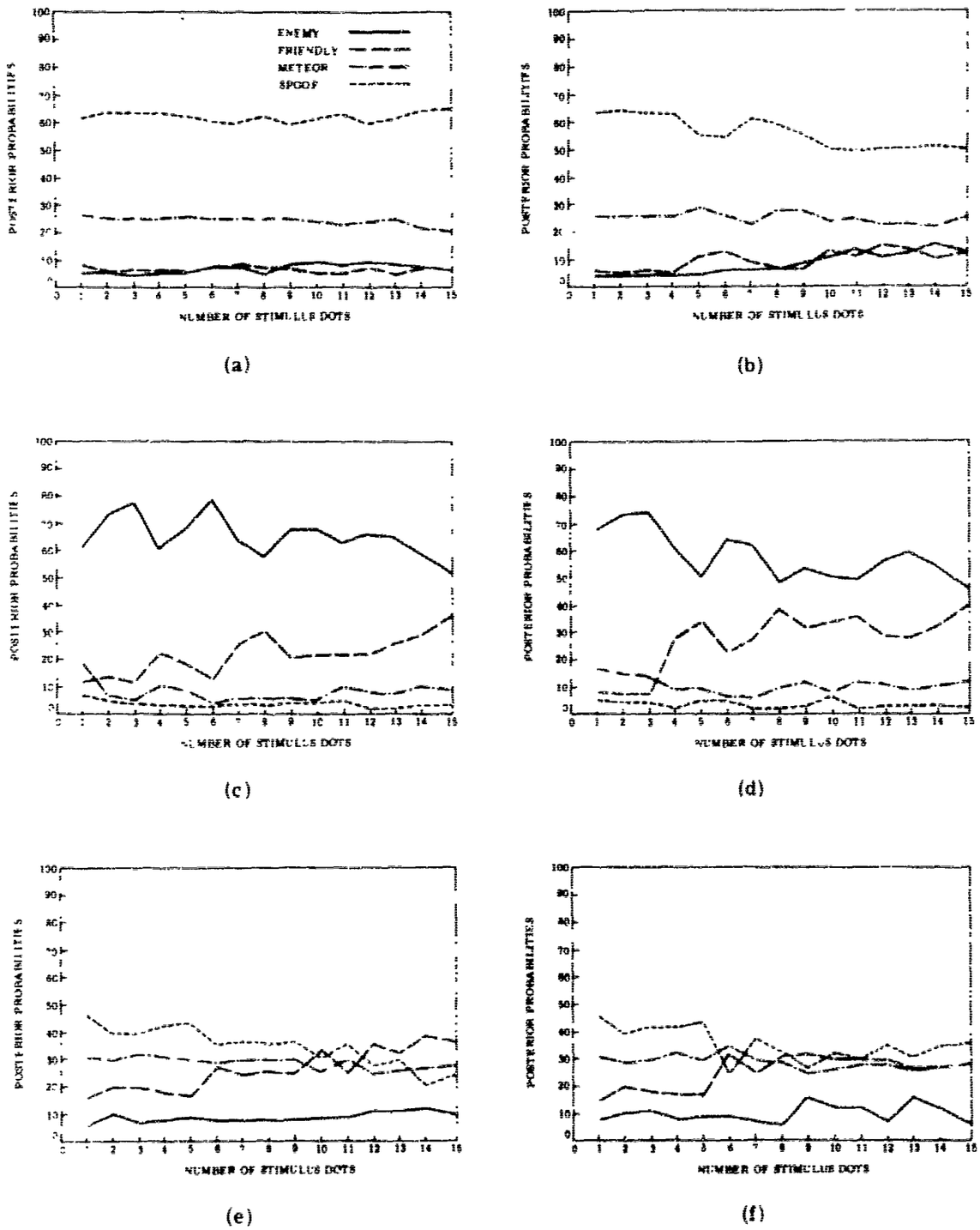


FIGURE 5. REPRESENTATIVE PLOTS OF SUBJECTS' NORMALIZED ESTIMATES. (a) Subject One, Sequence 28, Ordered. (b) Subject One, Sequence 28, Scrambled. (c) Subject Five, Sequence 12, Ordered. (d) Subject Five, Sequence 12, Scrambled. (e) Subject Two, Sequence 38, Ordered. (f) Subject Two, Sequence 38, Scrambled.

theorem would permit them to be, and stopped modifying their opinions in the light of additional information while they were still very far from posterior probabilities of one and zero.

2.1.2.4. Scatterplots. To determine whether this conservatism was consistent, scatterplots were constructed of the normalized posterior probabilities estimated by each subject as a function of Bayesian posterior probabilities for the ordered presentations. In Fig. 6 one set of scatterplots is shown for Subject One, who was neither the most Bayesian nor least Bayesian subject. Two variables have been retained. One is the number of stimulus dots. This variable is represented at a different value in each row. The first row is for one dot, the second for three, the third for six, and the fourth for nine. Because Bayesian probabilities (though not subjects' estimates) generally go to zero or one for more than nine dots, no further plots were made. The other variable is prior probability of the Enemy hypothesis. The first column is for all sequences with a prior probability of 0.10, the second for 0.25, and the third for 0.67. Estimates for the individual hypotheses have not been distinguished on these plots because more detailed analysis showed nothing systematically meaningful, except for the occasional underestimation of the Enemy hypothesis.

Subject One showed remarkably Bayesian performance for one dot. He seems to have used the prior probabilities effectively and to have been able to modify them properly on the basis of the first dot. But he became progressively less Bayesian as he obtained more information. His deviations from Bayes's theorem, however, were relatively consistent. He appears increasingly to have underestimated high posterior probabilities and overestimated low ones, until by the ninth slide the best fitting lines through his scatterplots would be almost horizontal.

Subject Three does not show such consistency. He initially underestimated the low posterior probabilities and overestimated the high probabilities. However, in general, the best fitting lines through his scatter plots would be nearly 45° lines. The other subjects showed varying degrees of consistency. The underestimation-overestimation tendencies of these subjects varied with the number of dots and were often confounded with prior probability.

These plots clearly show no single function relating their posterior estimates to Bayesian posterior probabilities for all subjects. Some subjects are more Bayesian near the beginning of the sequence, others nearer the end, this depends, in part, on the prior probability of the sequence. The variable that has the most pronounced effect on the relationship between posterior estimates and Bayes's probabilities is the number of stimulus dots.

2.1.2.5. A Performance Index. In order to show, on only one plot, the total performance of an individual subject, we devised a Performance Index (PI). Squared deviations from Bayes's theorem are misleading indices of performance. If a very conservative subject simply set the

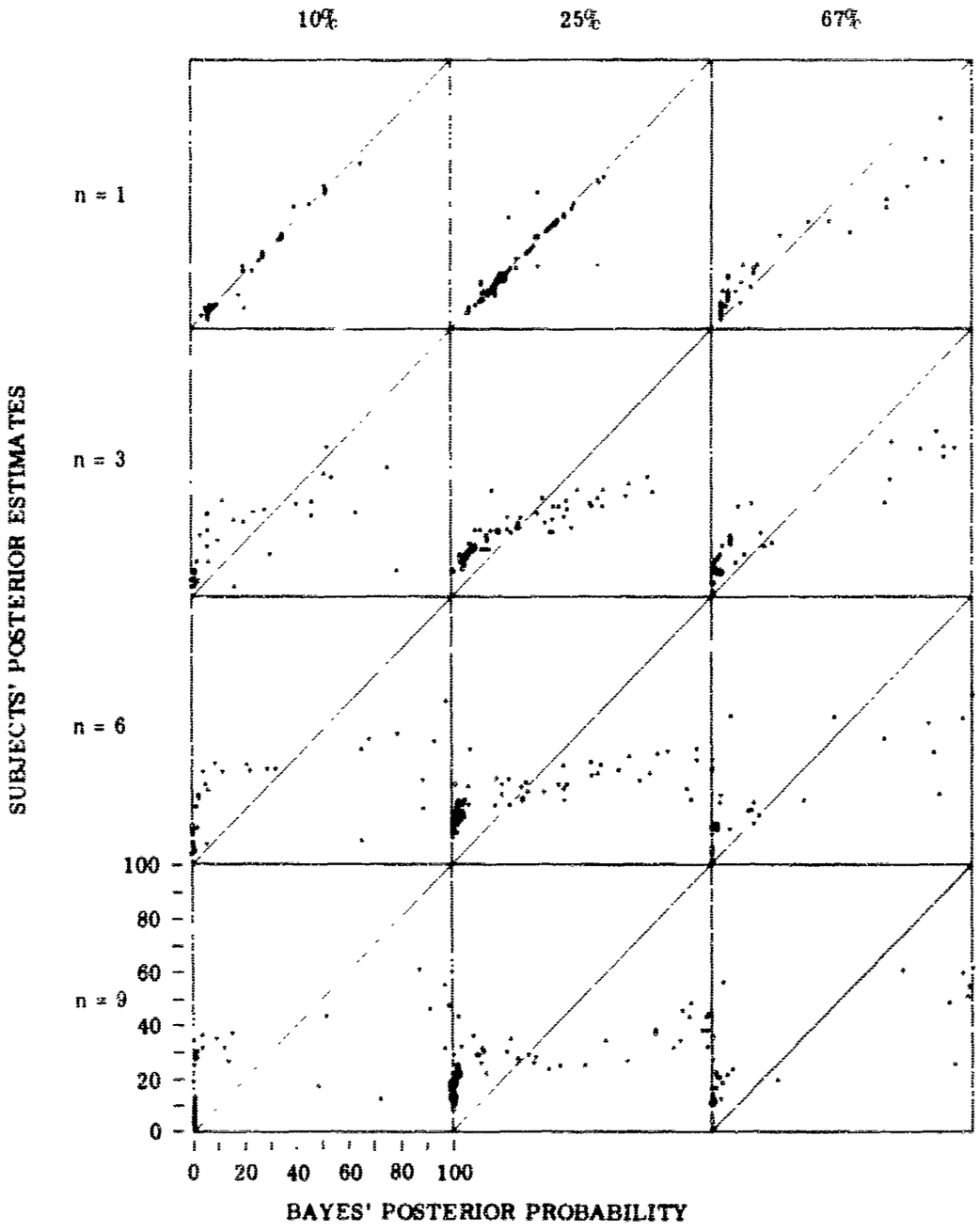


FIGURE 6. SCATTERPLOTS OF BAYES'S POSTERIOR PROBABILITY AS A FUNCTION OF POSTERIOR ESTIMATES BY SUBJECT ONE. Each row represents a different value of the number of stimulus dots. The three plots in each row are for those ordered sequences with enemy prior probability of 10%, 25%, and 67%.

posterior estimation levers at 0.25 regardless of the stimulus, his squared deviations would be lower for the more ambiguous, and thus presumably more difficult, sequences such as 38. And his squared deviations would be higher for the less ambiguous, easier, sequences such as 28. While it is obvious that this subject's performance is more like Bayes's theorem for the more ambiguous sequences than for the less ambiguous ones, it would be misleading to conclude that the quality of the subject's performance is different when he deals with the ambiguous ones than when he deals with the unambiguous.

Thus, a good Performance Index should have the properties of indicating very non-Bayesian performance and remaining constant with varying number of dots whenever subjects leave their levers at 0.25. It should also indicate perfect Bayesian performance and remain constant whenever subjects make estimates identical to numbers calculated according to Bayes's theorem. A ratio of squared deviation scores will exhibit these properties:

$$PI_n = \frac{\sum_i [\psi_n(H_i|D) - P_n(H_i|D)]^2}{\sum_i [0.25 - P_n(H_i|D)]^2} \times 100$$

where $\psi_n(H_i|D)$ is the normalized posterior probability of hypothesis H_i estimated by a given subject for a given number of dots, and $P_n(H_i|D)$ is the posterior probability of hypothesis H_i from Bayes's theorem for a given number of dots. In words, this measure is defined as the ratio of the sum over the four hypotheses of the squared deviations of an individual subject's posterior estimates from Bayes's theorem to the sum over the four hypotheses of the squared deviations of 0.25 from Bayes's theorem.

If a subject is perfectly Bayesian, his PI will be zero. If he leaves his levers at 0.25, his PI will be 100; 100 is therefore a kind of baseline or definition of absurdly poor performance. But if a subject gets a score of 100, he did not necessarily have all his levers at 0.25; he only indicated settings that gave summed deviations precisely the same as those set at 0.25. One difficulty with this measure is that only the values 0 and 100 are easily interpretable. Figure 7 shows PI as a function of the number of dots averaged over all sequences (ordered and scrambled) with the same prior probability. In interpreting these plots it is necessary to keep in mind one particular property of Bayes's theorem as more and more data are collected, the prior probability becomes more and more irrelevant to the value of the posterior probability. This is illustrated in Fig. 8, for Sequence 41. Enemy and Spoof probabilities are plotted (Friendly and Meteor probabilities are very low) for Enemy and Spoof prior probabilities of 0.67 and 0.11, respectively, and for 0.25 and 0.25. For more than five dots, the curves are reasonably close to one another. This is even more marked in sequences where the probabilities do not cross and

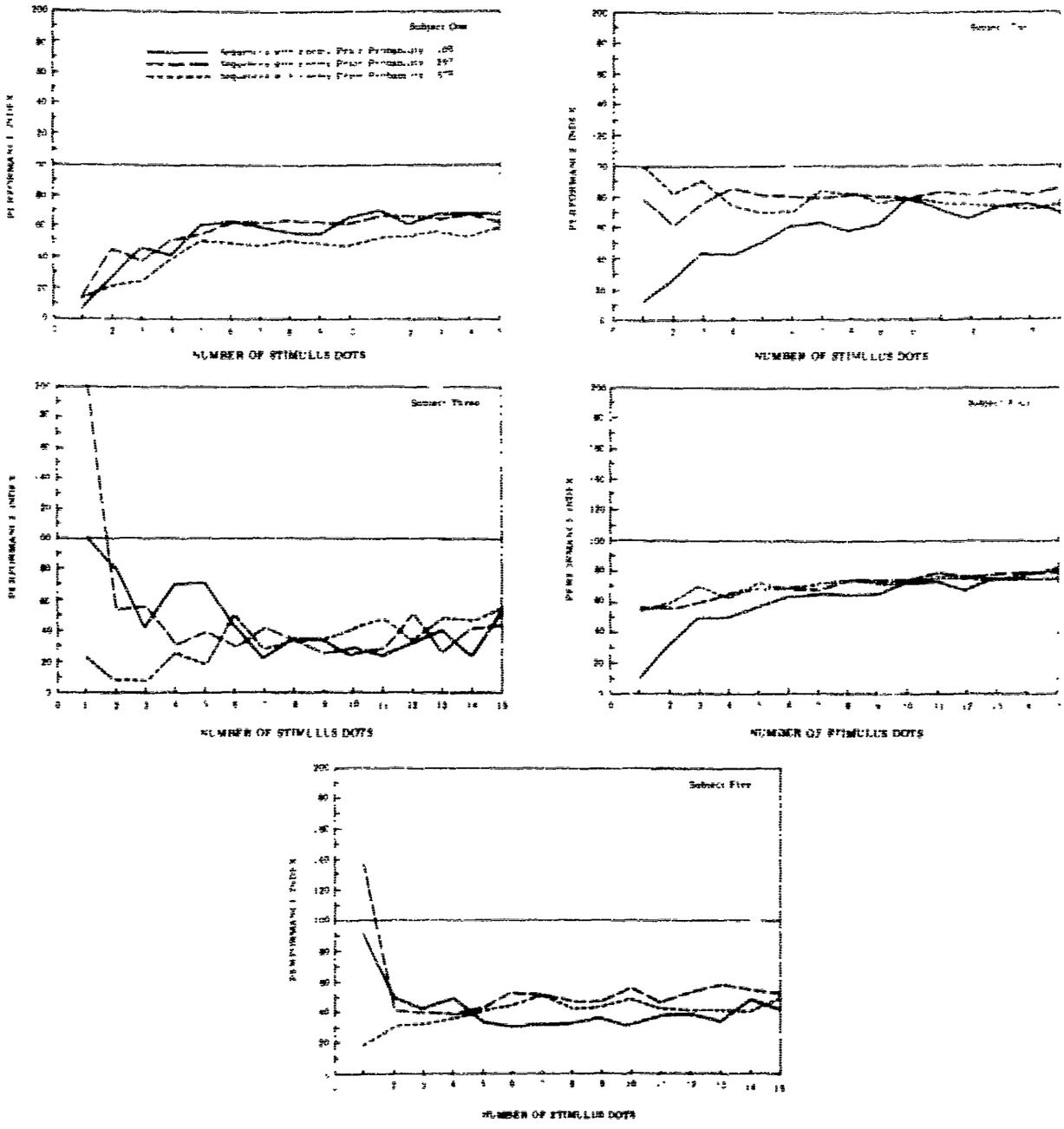


FIGURE 7. PERFORMANCE INDEX AS A FUNCTION OF NUMBER OF STIMULUS DOTS. Each curve represents performance on all sequences with the same prior probabilities.

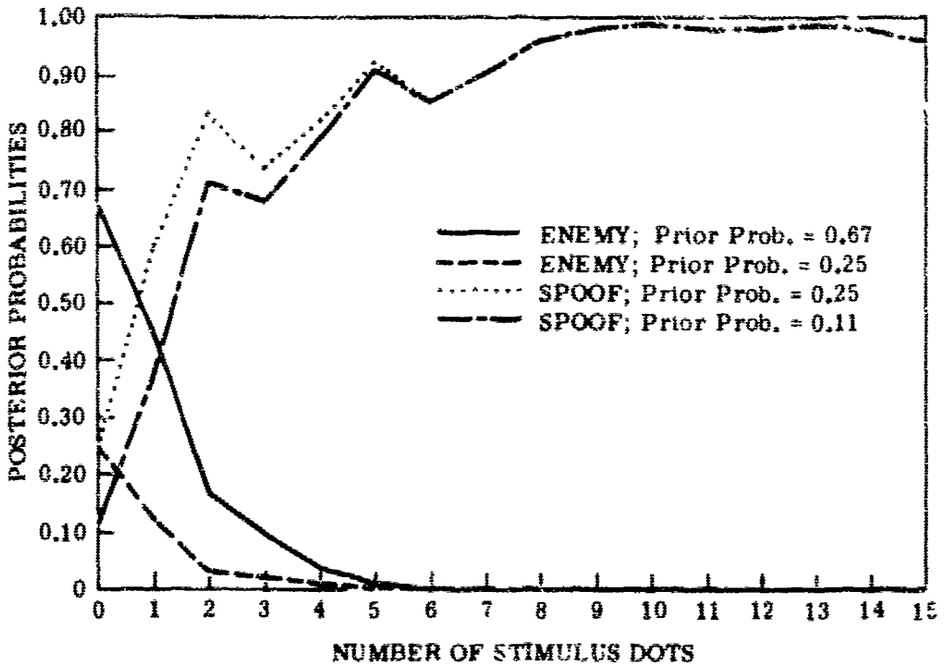


FIGURE 8. BAYESIAN POSTERIOR PROBABILITIES AS A FUNCTION OF NUMBER OF STIMULUS DOTS FOR SEQUENCE 41

where they quickly converge on 0 or 100. Thus, for a subject to be perfectly Bayesian, he would have to give less weight to the prior probability the more dots he sees. Failure to correctly weight the prior information relative to the observed data will cause the PI to be greater than zero.

The Performance Index can give only a rough indication of why subjects deviate from Bayesian performance. Some subjects show PI scores that are initially very low (close to Bayes's theorem) and then increase to a constant value. Others start very high, sometimes above 100, but then decrease to a constant value. Notice that the constant value attained by these latter subjects is usually lower than that of the subjects who start low.

A PI curve that starts low and then increases can be explained as characteristic of performance which tends to weight the prior information too heavily, at least for n greater than one. As more and more dots appear, giving too much weight to the prior probability will result in a gradually increasing PI. A curve that starts high and then decreases would result from performance that tends to weight the prior information insufficiently; as the data accumulate, ignoring the prior probability becomes less and less serious, and the PI decreases. In both cases, a constant value is reached because, on the average, the performance of neither the subject nor Bayes's theorem changes very much after about seven dots. The constant value should be lower (better) for those subjects who weight prior information less heavily than those who do the opposite, since the prior information becomes increasingly irrelevant as data increases in amount. This relative difference in the constant values will only be true, of course, if the differences in estimating the conditional probabilities are not too great.

Thus, Subjects One and Four, and to a lesser extent, Two, appear to weight prior information too heavily. Note that the shape and smoothness of the curves for Subject One agree very well with what can be predicted from his scatterplots. Subjects Three and Five appear to underweight prior information, at least for sequences with prior probabilities of 0.10 and 0.25. And their constant values are less than those for Subjects One, Two, and Four. The data of Subjects One and Four suggest that high prior probabilities may partially correct the tendency to underweight prior information, for the shapes of their curves for sequences whose prior probability is 0.67 are quite different from changes produced by the other prior probabilities.

The terminal value of the performance index for Subject Three — a little less then 40 — is smaller than that for any other subject, but not much. The central tendency of his performance is closer to Bayes's theorem than that of any other subject, but his estimates scatter more widely around the central tendency than do those of any other subject. This observation highlights a deficiency of the Performance Index (and of any other error-measure based on mean squared error); it cannot discriminate between random and constant error. The errors found in this experiment are mostly constant rather than random errors — as is often the case when performance is being compared with some standard of perfection.

Thus, the PI plots show that subjects' deviations from Bayes's theorem can be partly explained by failure to weight the prior information properly. And they confirm that actual level of performance is dependent on subjects, prior probabilities, and the number of dots. No further information was gained by plotting PI as a function of the number of dots for all sequences in the same set. PI plots averaged over subjects for the interaction between prior probability and sets showed nothing of interest.

2.1.3. DISCUSSION. The primary conclusion indicated by this experiment should surprise no one: men are suboptimal processors of probabilistic information. Several things about the finding are a little surprising. For one thing, the size of the discrepancy is large—surprisingly large compared with our expectation. For another thing, we have failed to find any subjects who consistently leaped to a conclusion more quickly than is justified by the evidence. In fact, most subjects simply refused to estimate an extremely large posterior probability at all, in spite of the fact that they seemed to find it easier to judge what diagnosis was favored by a new item of information than to judge what diagnosis was made less probable by that item. Even in college populations, some men must have a tendency to jump to conclusions; yet this experiment has failed to exhibit any such tendency in any subject. Perhaps such men do not find service as paid subjects in psychological experiments sufficiently attractive to volunteer for it.

Underestimation of high probabilities and overestimation of low ones, often reported in decision-theory experiments (e.g., Mosteller and Nogee [11], Preston and Baratta [12]), are not invariably found in this experiment. They are largely absent in one subject, dependent on amount of information for others, and confounded with prior probability for all. Still, the congruence between the findings of this experiment and those of experiments concerned with estimation of relative frequencies (Philip [1], Stevens and Galanter [2]) or with probabilities inferred from choices among bets (Griffith [13], Mosteller and Nogee [11]) suggests an underlying tendency toward conservatism in estimation and use of probabilities over a wide class of tasks—at least among college students. (But for conflicting evidence see Dale [14], and for an argument that things are more complicated than this, see Edwards [8].)

Other factors that may have influenced performance are display parameters. A pretest of several different methods of displaying P(D|H) resulted in the displays shown in Fig. 1. Although no subject complained about these displays, the question lingers whether they may have accounted, in part, for the conservative behavior. None of the displays shows a sector probability greater than 0.25. Perhaps the (necessarily) low numbers on the P(D|H) displays suggested to the subjects that there should not be too much difference between their estimates, whatever the data.

Further, we assumed that it was necessary to display only Enemy prior probability because subjects would distribute the remaining probability equally among the other three

hypotheses. In view of the finding that subjects' estimates do not always sum to one, it is questionable whether they knew how much probability was left to be distributed among the remaining alternatives. The consistency shown by some subjects on the scatterplots indicates that this is probably not a serious problem, but prior probabilities will be displayed for each hypothesis in future experiments.

Another methodological issue concerns the utilities that some subjects may have attached to the particular hypotheses. One subject showed occasional underestimation of Enemy probabilities, suggesting that he was especially conservative in making this diagnosis. However, the scatterplots were originally drawn so that estimates for each hypothesis could be examined separately, and the estimates made for one hypothesis very rarely showed any consistent deviation from the other estimates. So this issue is probably not very important either.

The remaining experiments reported in this section examine two factors that could have contributed to the conservatism of the subjects. One is the artificiality of the stimulus display, and the other concerns the method of responding.

2.2. EXPERIMENT TWO[3]

In Experiment One the stimulus dots were constrained to appear in only three of the twelve sectors of the display. If each sequence of 15 dots had been randomly generated under the truth of exactly one of the hypotheses, then the dots would have been distributed over more than three sectors. The artificial constraint on the distribution of dots produced sequences that looked unlike any of the hypotheses; that might be why subjects estimated conservatively.

This experiment tests the hypothesis that conservative posterior probability estimation in the original experiment was due, at least in part, to the artificial constraint on the dots variable. For the experiment, new sequences were generated, each having posterior probabilities approximately equal to the posterior probabilities of a sequence in the original study; however, the dots were distributed over several sectors, to look like a more representative sample than did the original sequences.

2.2.1. METHOD

2.2.1.1. Apparatus. Conditional probability displays and apparatus were the same as those used in Experiment One.

2.2.1.2. Stimuli. Subjects were shown eight sequences of fifteen dots each. For all sequences, prior probabilities were given as 0.25. Four of them were sequences 35, 44, 24, and

[3] This experiment was run by Richard Norman.

41 of Experiment One, with the dots appearing in only three of the twelve sectors. Four new sequences were constructed in which the dots were distributed over more than three sectors. However, the new sequences had posterior probabilities very nearly identical to the posterior probabilities for the old sequences when the posterior probabilities for all sequences were calculated from Bayes's theorem using prior probabilities of 25%. By the fifteenth dot, the posterior probabilities of every sequence are near one or zero. Figure 9 shows the distributions of dots for the original sequence 44, and for its equivalent new sequence, Fig. 10 shows the Bayesian posterior probabilities for these sequences.

2.2.1.3. Procedure. Each subject was shown all eight sequences in random order, in two sessions lasting a total of about three hours. Conditions were comparable to the ordered sequence presentation of the Phillips, Hays, and Edwards experiment.

2.2.1.4. Subjects. Four men, University of Michigan undergraduates, served as subjects. They were paid $1 5 per hour.

2.2.2. RESULTS. The amount that subjects revised their estimates from one dot to the next was generally more conservative then the revision of probability calculated from Bayes's theorem.

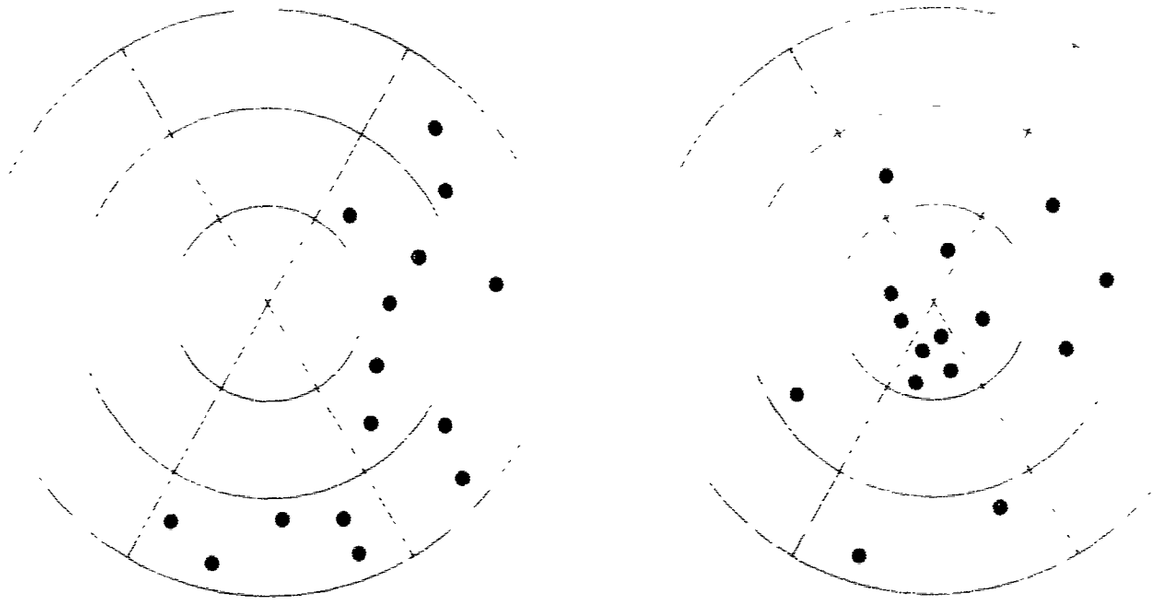
An analysis of variance was computed using as the dependent variable the absolute deviations of subjects' estimates from Bayes's theorem for the correct hypothesis. The fifteen deviations generated from a single sequence by one subject were treated as independent measures, an assumption justified by the insignificance of the order-of-presentation variable in Experiment One.

Three independent variables were examined: (1) Distribution of dots, representative or unrepresentative; (2) Sequences; and (3) Subjects. The sequences variable is, of course, nested within the dots variable. Table I shows the variance due to subjects is highly significant. Variance due to dots is mildly significant, while the dots-by-subjects interactions is not significant.

TABLE I. ANALYSIS OF VARIANCE OF SUBJECTS' DEVIATIONS FROM BAYES'S THEOREM

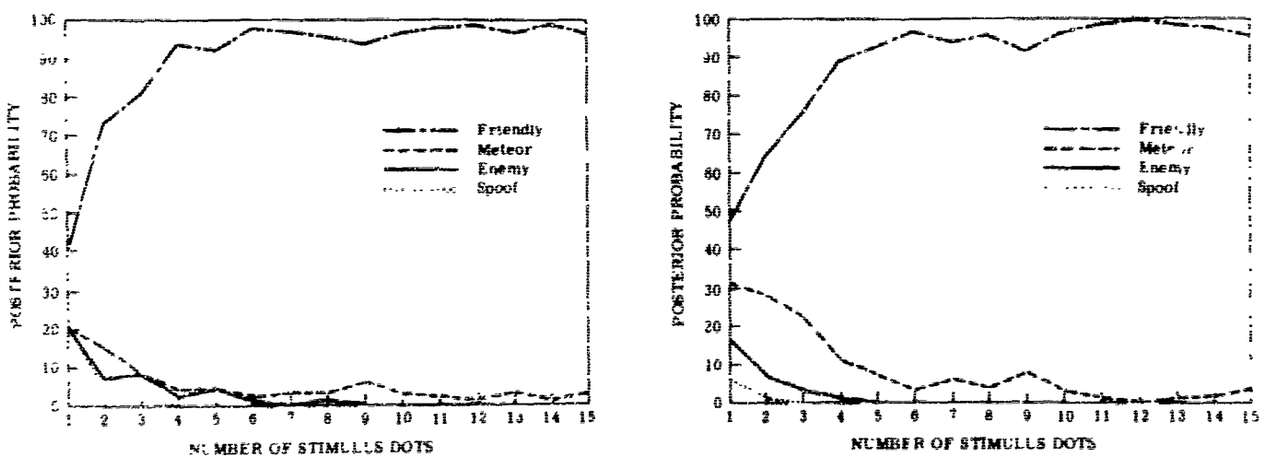
Source	df	MS	F
Dots (D)	1	1,491.08	4.87*
Sequences (Se) nested in D	6	532.61	
Subjects (Ss)	3	10,654.42	34.81**
D x Ss	3	395.56	1.29
Se (D) x Ss	18	1,194.38	
Within cell	480	269.96	
Pooled error	504	306.11	

*P < .05
**P < .01



(a) (b)

FIGURE 9. DISTRIBUTION OF STIMULUS DOTS FOR OLD AND NEW SEQUENCES. (a) Distribution of stimulus dots for Old Sequence 44. (b) Distribution of stimulus dots for New Sequence 44.



(a) (b)

FIGURE 10. BAYESIAN POSTERIOR PROBABILITIES FOR OLD AND NEW SEQUENCES. (a) Old Sequence 44. (b) New Sequence 44.

2.2.3. DISCUSSION. Once again, the primary finding is conservatism. Every subject in this experiment extracted less certainty from the data than is justified by the Bayesian calculations.

The analysis of variance indicates that the distribution of dots does make some difference, though a glance at the mean square of the dots variable shows that this variable is not a large source of variation. The magnitude of the deviation scores is shown in Table II. The total deviation score for the old sequences indicates that performance was more Bayesian for the old sequences. Thus, the distribution of dots certainly does not explain the conservatism at all; what effect the dots variable does have seems to operate in a direction opposite to that hypothesized.

TABLE II. DEVIATIONS FROM BAYES'S THEOREM OF SUBJECTS' ESTIMATES ON THE CORRECT HYPOTHESIS

	Sequences	
Subject	New	Old
1	2,287	1,995
2	2,428	2,324
3	2,086	2,110
4	1,338	864
Total	8,139	7,293

Deviations are summed over the fifteen estimates per sequence and over four sequences of each class.

To satisfy ourselves that the new sequences had posterior probabilities nearly equal to the corresponding old sequences, we computed an analysis of the variance of the Bayesian posterior probabilities. The results, shown in Table III, do indeed confirm that the differences between old and new sequences are very small. Thus, interpretation of the first analysis of variance is not contaminated by differences in old and new sequences.

TABLE III. ANALYSIS OF VARIANCE OF POSTERIOR PROBABILITIES FOR OLD AND NEW SEQUENCES, CALCULATED FROM BAYES'S THEOREM

Source	df	MS	F
Dots (D)	1	216.01	n.s.
Sequences nested in D	6	1,694.26	4.32**
Within cell	112	388.12	

** $P < .01$

We conclude, then, that conservatism in this task is unaffected by the representative character of the stimulus display.

2.3. EXPERIMENT THREE[4]

Experiment One suggested that there is a correlation between task difficulty and the degree to which subjects approach Bayesian performance in processing probabilistic information. Many subjects appeared to be more Bayesian when the sequences were simple, that is, when the data clearly pointed to only one hypothesis as the most likely one. Sequences appeared to become more difficult as the information became more ambiguous and contradictory. The experiment we are now reporting studies this variable, using three sequences representing three levels of difficulty, and one other variable.

The other variable studied concerns the difference between the two possible interpretations of Bayes's theorem. The input to the theorem can be correctly expressed in two ways. One way is to use the conditional probability of n dots for each hypothesis with the prior probability for n = 1. This we will call the nonsequential version of Bayes's theorem. The other way is to use the conditional probabilities of only the new dot at slide n with prior probabilities that are the posterior probabilities from slide n - 1. This we will call the sequential model. Both methods of calculation lead to the same posterior probabilities.

Since subjects in the original experiment were presented with dots that accumulated, and were required to reset their levers after each set of estimates, the cards were stacked in favor of their adopting a nonsequential mode of behavior, though not necessarily Bayesian. The present experiment examines the effects of presenting only one dot on the viewing screen for any value of n, where n is the total number of dots shown. Subjects were not required to reset their levers. In fact, they were told to revise on trial n + 1 the settings they left at trial n. In other words, they were encouraged to use their posterior settings at trial n - 1 as the prior probabilities for trial n.

Thus, the question of interest is whether subjects are more or less Bayesian for the sequential mode than for the nonsequential mode.

2.3.2. METHOD

2.3.2.1. Subjects. Six summer students were subjects. All were volunteers hired through the Student Employment Office, and each was paid $1.25 per hour. All subjects completed the experiment in less than two hours.

[4] This material was prepared by Lawrence D. Phillips and Ward Edwards.

2.3.2.2. Apparatus and Method. Apparatus and conditional probability displays were the same as used in Experiment One. Each subject was presented first with three sequences from the original study, numbers 12, 28, and 38 (see Fig. 3). Sequence 28 is relatively easy (the data clearly indicate only one hypothesis as the "correct" one), 12 is moderately difficult (the data point ambiguously to two hypotheses), and 38 is difficult (the data are ambiguous about all four hypotheses). The order in which these sequences were presented was completely counterbalanced for the six subjects. The prior probabilities were displayed above each of the conditional probability displays and remained in view throughout the entire sequence of fifteen dots. Response levers had to be reset after each slide.

Following these sequences, the subjects were presented with the first three sequences, the only differences being that the data and the conditional probability displays were inverted and reversed and dots did not accumulate. These sequences were designated 62, 78, and 88 (add 50 to the original sequence number) and were presented to each subject in the same order in which the first three were given. The prior probabilities were displayed on a slide just prior to the first dot. The subjects were required to set their levers according to the prior probabilities displayed on the first slide, and were told to revise that estimate when shown the first dot. They were not allowed to reset their levers to zero, and were instructed to revise their lever settings as they received new information.

Normalization of posterior estimates was required under both presentation conditions. The cover story attempted to attach equal utilities to the four hypotheses.

Subjects were asked at the completion of all sequences if they noticed any similarities between the first three sequences and the latter three. No subject reported that he did.

2.3.3. RESULTS. On Sequences 28 and 78 all subjects tended to underestimate the high probabilities and overestimate the probabilities for the other three hypotheses. This tendency is evident to a lesser degree in Sequences 12 and 62, but not very apparent in Sequences 38 and 88; this is probably because the Bayesian posterior probabilities are not as extreme for these sequences.

Performance Indices were computed for each subject on each sequence for each value of n. An analysis of variance on these PI's gave the results shown in Table IV. Because only one observation appeared in each cell, the error term used in the analysis was the figure representing the mean squares of the sequences times presentation times dots times subjects variable.

In interpreting this analysis of variance, it is important to keep in mind that PI is being examined, so the experimental variables must be understood to affect the degree to which subjects were successful in approaching Bayesian performance. Three of the main effects are significant. Sequences, for one: subjects are less Bayesian for the more difficult sequences.

The number of dots is mildly significant, the trend being towards the less Bayesian performance as the number increases, though there are some exceptions for some subjects at some sequences. For Sequences 28 and 78 all subjects became less Bayesian with more dots; for the other two sequences the PI peaks sharply and rather irregularly, but there is enough consistency among subjects to give a sequences x dots interaction.

TABLE IV. SUMMARY OF ANALYSIS OF VARIANCE

Source	df	MS	F
Sequences	2	104,525.5	43.79**
Presentation (sequential or nonsequential)	1	5,314.0	2.23†
Number of dots	14	5,111.1	2.14*
Subjects	5	17,870.4	7.49**
Sequences x presentation	2	2,554.0	1.07†
Sequences x dots	28	9,729.1	4.08**
Sequences x subjects	10	13,835.1	5.80**
Presentation x dots	14	1,457.3	-†
Presentation x subjects	5	8,913.2	3.73**
Dots x subjects	70	4,167.5	1.75**
Sequences x presentation x dots	28	2,017.3	-†
Sequences x presentation x subjects	10	9,048.8	3.79**
Sequences x dots x subjects	140	3,334 3	1.40*
Presentation x dots x subjects	70	2,107.1	-†
Sequences x presentation x dots x Ss	140	2,386 8	-

** $P < .01$
* $P < .05$
† n.s.

Individual differences among subjects are high; thus, the subjects main effect is significant. For some subjects the method of presentation makes a difference. This is not true for all subjects, however, so there is a presentation x subjects interaction, but not a main effect due to presentation. Further, for those to whom presentation condition does make a difference,

2.3.2.2. Apparatus and Method. Apparatus and conditional probability displays were the same as used in Experiment One. Each subject was presented first with three sequences from the original study, numbers 12, 28, and 38 (see Fig. 3). Sequence 28 is relatively easy (the data clearly indicate only one hypothesis as the "correct" one), 12 is moderately difficult (the data point ambiguously to two hypotheses), and 38 is difficult (the data are ambiguous about all four hypotheses). The order in which these sequences were presented was completely counterbalanced for the six subjects. The prior probabilities were displayed above each of the conditional probability displays and remained in view throughout the entire sequence of fifteen dots. Response levers had to be reset after each slide.

Following these sequences, the subjects were presented with the first three sequences, the only differences being that the data and the conditional probability displays were inverted and reversed and dots did not accumulate. These sequences were designated 62, 78, and 88 (add 50 to the original sequence number) and were presented to each subject in the same order in which the first three were given. The prior probabilities were displayed on a slide just prior to the first dot. The subjects were required to set their levers according to the prior probabilities displayed on the first slide, and were told to revise that estimate when shown the first dot. They were not allowed to reset their levers to zero, and were instructed to revise their lever settings as they received new information.

Normalization of posterior estimates was required under both presentation conditions. The cover story attempted to attach equal utilities to the four hypotheses.

Subjects were asked at the completion of all sequences if they noticed any similarities between the first three sequences and the latter three. No subject reported that he did.

2.3.3. RESULTS. On Sequences 28 and 78 all subjects tended to underestimate the high probabilities and overestimate the probabilities for the other three hypotheses. This tendency is evident to a lesser degree in Sequences 12 and 62, but not very apparent in Sequences 38 and 88; this is probably because the Bayesian posterior probabilities are not as extreme for these sequences.

Performance indices were computed for each subject on each sequence for each value of n. An analysis of variance on these PI's gave the results shown in Table IV. Because only one observation appeared in each cell, the error term used in the analysis was the figure representing the mean squares of the sequences times presentation times dots times subjects variable.

In interpreting this analysis of variance, it is important to keep in mind that PI is being examined, so the experimental variables must be understood to affect the degree to which subjects were successful in approaching Bayesian performance. Three of the main effects are significant. Sequences, for one, subjects are less Bayesian for the more difficult sequences.

The number of dots is mildly significant, the trend being towards the less Bayesian performance as the number increases, though there are some exceptions for some subjects at some sequences. For Sequences 28 and 78 all subjects became less Bayesian with more dots; for the other two sequences the PI peaks sharply and rather irregularly, but there is enough consistency among subjects to give a sequences x dots interaction.

TABLE IV. SUMMARY OF ANALYSIS OF VARIANCE

Source	df	MS	F
Sequences	2	104,525.5	43.79**
Presentation (sequential or nonsequential)	1	5,314.0	2.23†
Number of dots	14	5,111.1	2.14*
Subjects	5	17,870.4	7.49**
Sequences x presentation	2	2,554.0	1.07†
Sequences x dots	28	9,729.1	4.08**
Sequences x subjects	10	13,835.1	5.80**
Presentation x dots	14	1,457.3	-†
Presentation x subjects	5	8,913.2	3.73**
Dots x subjects	70	4,167.5	1.75**
Sequences x presentation x dots	28	2,017.3	-†
Sequences x presentation x subjects	10	9,048.8	3.79**
Sequences x dots x subjects	140	3,334.3	1.40*
Presentation x dots x subjects	70	2,107.1	-†
Sequences x presentation x dots x Ss	140	2,386 8	-

**P < .01
*P < .05
† n.s.

Individual differences among subjects are high: thus, the subjects main effect is significant. For some subjects the method of presentation makes a difference. This is not true for all subjects, however, so there is a presentation x subjects interaction, but not a main effect due to presentation. Further, for those to whom presentation condition does make a difference,

the direction of this difference varies according to sequence. The effect of subjects is apparently strong, for there is no significant sequences x presentation interaction (although lack of this interaction may be due to the low number of degrees of freedom).

Sequences x subjects is significant; while some subjects are most nearly Bayesian on the easiest sequence and least Bayesian on the hardest, some are not. The significance of the triple interaction, sequences x subjects x dots, undoubtedly results only from the great number of degrees of freedom.

Some subjects tend to be less Bayesian at the start of a sequence and more Bayesian near the end, while others reverse this trend. This leads to the dots x subjects interaction.

To summarize, the only highly significant main effect is that of sequences. This means that highly conflicting, ambiguous information leads to performance which is less Bayesian than that produced by unambiguous information. Other factors also influence performance, but are less important.

2.3.4. DISCUSSION. This experiment shows that more ambiguous information produces less Bayesian performance. It seems likely that ambiguity interacts with the number of hypotheses considered by the subject. Of course the subject, being conservative, may be considering as plausible hypotheses that have negligible Bayesian posterior probability; it may be possible to improve performance in multihypothesis situations by reducing the number of hypotheses under active consideration as rapidly as the data permit.

The other major finding of the experiment is that sequential vs. nonsequential presentation of data makes very little difference. This finding is not too surprising. Experiment One showed that subjects were treating each slide as a separate problem, whether or not it appeared in ordered sequence. In this experiment, no subject performed better under sequential than under nonsequential conditions of presentation; some performed worse. Apparently the difference in the kind of information processing required makes little difference to performance. Of course all information necessary to calculate valid posterior probabilities is present under both conditions. If, in the sequential (only one dot on the screen at a time) mode of presentation the subject had been required to reset his estimation levers to zero, thus putting a load on his memory, presumably performance would have deteriorated.

Methodological issues cloud the picture. All subjects were first presented with the three sequences in which displayed dots accumulate, and then with those wherein the dots appear sequentially. This order may have caused subjects to try to perform the sequential task in the same manner as the nonsequential even though their instructions for the former were to revise their last lever setting as they gained new information. Since subjects were not told

that the revision was to be based only on the new dot, they could have based their revision on the new dot and on what they remembered of the previous dots. This suggests the possibility that memory factors are affecting performance of the sequential task, and that it is for this reason that performance there doesn't consistently differ from performance on the nonsequential problem.

2.4. OVERALL DISCUSSION

Experiments Two and Three suggest that conservatism is a very pervasive phenomenon, little affected by different stimulus displays or different response modes. This conservatism in processing information conforms to our intuition and to our observations. We believe that men typically want to be more certain than they should want to be, and seek too much information: that generalization combined with the rules of the game is often enough to play winning poker. Furthermore, intuition suggests an interaction with payoff: the larger the payoff, the larger the excess of information that a decision-maker seeks over what he should seek. Anecdotal observations that people seek too much information have often been attributed to a "desire for certainty," or to a "dislike of intermediate probabilities," or to a "fear of failure in excess of desire for success," or to some similar motivational construct. These findings suggest a different interpretation: people seek too much information not because they want too much certainty, but rather because they cannot extract from the information they have as much certainty as it in principle justifies. In other words, the suboptimal behavior may be the result of intellectual, not motivational, deficiencies.

Two speculations about the reason for the intellectual deficiencies that lead to conservatism in information processing occur to us. First, the real world is always changing; certain kinds of hypotheses that seem true today may not be true tommorrow. Thus, evidence about the truth of one hypothesis in the real world may be misleading, not because the hypothesis was not true at the time the evidence was collected, but rather because the world has changed since then. One possible defense against being misled is to resist persuasion, to require large amounts of evidence before acting. It is not difficult to imagine a learning process for acquiring that defense: experiences should not be hard to come by in which acting in accord with the weight of the evidence and being wrong leads to punishment.

A second similarly speculative explanation of the conservatism concerns the dependence of data. If two data are independent given a hypothesis, then

$$P(D_j|H) = P(D_j|H, D_k) \text{ and } P(D_k|H) = P(D_k|H, D_j)$$

for that hypothesis under consideration. (Note that the relation of independence is a relationship among at least two data and a hypothesis, so that data may be independent given one hypothesis and

dependent given another.) In the real world, data are often not independent of one another. Moreover, one kind of dependence is far more frequent than any other: repeated observations of the same datum. Thus when you look around your office and see John there, and a moment later look again and again see John there, you do not conclude that there are two people in your office; instead you conclude that John has remained there. Men may be accustomed to discount the significance of items of evidence that resemble one another. If so, one might expect that qualitatively different items of evidence would have more impact on opinion than qualitatively similar items.

In any case, our findings strongly suggest that men should not be required to estimate posterior probabilities in information-processing systems. If the conservatism in information processing suggested by this experiment is also reflected in decision-making, questions are raised about the quality of men's decisions in such cases.

3
THE EFFECT OF A FLATTENED CONDITIONAL PROBABILITY DISTRIBUTION ON PROBABILITY ESTIMATION[5]

In experimental situations where subjects are given a set of hypotheses, prior probabilities for the hypotheses, and conditional probability distributions for information or data given the respective hypotheses, the usual finding has been conservatism; subjects change their probability estimates less than the amount prescribed by Bayes's theorem.

A series of studies conducted by Harold C. A. Dale (1962, unpublished) approached the question of probability estimation as a training problem in probabilistic diagnosis. In the Dale studies, the subject is placed in a simulated war game. He is told that enemy forces may launch any one of four types of attack and his task is to estimate the probability of each as he is presented with a sequence of information concerning enemy activity. For each datum, four different values of P(D|H) are possible, one for each hypothesis; these values are displayed to the subject. Thus, this task is very similar to the one reported in Section 2. Here, too, the normative solution is given by Bayes's theorem.

Again, subjects were found to estimate conservatively. Several possible explanations for their conservatism were considered and examined by Dale. If subjects, rather than accepting the displayed conditional probability, operated with a conditional probability matrix that was flatter (having less variance than the objective display) then the outcome would be the observed conservatism. If, on the other hand, subjects did not employ the Bayesian multiplication rule but rather used some sort of addition of probabilities, conservatism would still prevail. A third possibility is that subjects, while accepting the multiplication rule, make consistent computational errors.

Studies of these possibilities indicate that subjects persist in conservatism even when allowed to set their own conditional probability distributions and prior probabilities. It also seems that to provide subjects with a demonstration of the multiplication rule and training in its use does not improve the accuracy of estimation unless subjects are allowed to actually carry out paper and pencil computation.

The persistence of conservatism led to conjecture as to whether there could be constructed a conditional probability matrix that would not result in conservatism; this question gave rise to the experiment reported here.

---

[5]This section was prepared by Melvin Guyer and Ward Edwards.

3.1 INTRODUCTION.

In a simulated war game, subjects were instructed to estimate the probabilities of each of four mutually exclusive hypotheses when provided with data and conditional probability displays for the data given the hypotheses, and an attempt was made to construct a set of conditional probability distributions that would lead subjects to revise their probability estimates more than the amount prescribed by Bayes's theorem. Achieving these results would suggest certain explanations of the conservatism found fairly consistently in similar situations reported in the literature.

3.2. METHOD

Each of 20 University of Michigan male undergraduate students was randomly assigned to one of four experimental groups in a 2 × 2-design experiment. Two sets of conditional probability matrices were devised. One, hereafter referred to as the "basic" matrix, had the form

	e_1	e_2	e_3	e_4	e_5
H_1	0.40	0.10	0.20	0.20	0.10
H_2	0.10	0.40	0.10	0.10	0.30
H_3	0.20	0.10	0.10	0.40	0.20
H_4	0.30	0.30	0.10	0.10	0.20

where H_1 through H_4 were a set of mutually exclusive hypotheses concerning the form of a possible enemy attack, and e_1 through e_5 were a set of possible messages that the subject might receive and whose impact on the probabilities of the hypotheses he would have to estimate. A second matrix, called the "degraded" matrix, was constructed by adding a constant of 2.00 to each value in the basic matrix and then normalizing. The degraded matrix had the following form:

	e_1	e_2	e_3	e_4	e_5
H_1	0.22	0.19	0.20	0.20	0.19
H_2	0.19	0.22	0.19	0.19	0.21
H_3	0.20	0.19	0.19	0.22	0.20
H_4	0.21	0.21	0.19	0.19	0.20

The labeling for the hypotheses and the messages was of course the same for both basic and degraded matrices.

The matrices were displayed to the subjects as sets of bar graphs, one for each hypothesis, on separate large sheets of white cardboard. Each graph was labeled so that the probability values could be read easily.

Additional apparatus included a "map" of a supposed enemy terrain with various strategic areas demarcated, and showing the location of an agent who would be the source of messages

concerning enemy activity. The subject was also given a chip board and 100 metal chips. The chip board was made up of four columns, each with ten troughs capable of holding ten chips. Each column of the board was labeled for one of the hypotheses and the number of chips placed in the columns by the subject indicated the subject's estimate of the probabilities.

Each of the subjects was run on both matrices, the assignment to order of matrices being random, and matrix order being an experimental treatment. Since each subject was to be run on both, it was necessary to construct two different sequences of ten messages each. The sequences not only had different orders of messages but also provided evidence for different hypotheses. The probability values of the respective hypotheses for each sequence were quite similar and the values at the end points of the sequences were almost identical. The assignment to sequence order was random, and was also an experimental treatment.

Each subject was seated before the map of enemy terrain with the conditional probability matrix displayed and the chip board close at hand. Initially the chips were distributed equally among the four columns and the subject was told that the present state of our knowledge concerning enemy activity justified this distribution of chips. The subject was instructed in the use of the chip board and was told the nature of the task. He was requested to make estimates of the probability of each hypothesis as messages from the agent came in (the messages were presented to the subject by the experimenter). The subject made his estimates and then redistributed the chips among the columns. The experimenter recorded the distribution of probabilities for the hypotheses after each message. After a subject had been run on the first sequence of messages he was given additional instructions to explain the introduction of the second matrix and was then run on the remaining sequence of messages.

3.3. RESULTS

Figure 11 shows the averaged subjective estimates, using the basic matrix, of the probability of the hypothesis confirmed by the data. The upper curve represents the Bayesian values of the posterior probabilities after each message is received. The middle curve represents the averaged scores for the group first run on the basic matrix; and the lower curve, the averaged scores for subjects run first on the degraded matrix and then on the basic. Of course the sequence is the same for all curves in Fig. 11.

Figure 12 gives the same information as Fig. 11, except that Sequence 2 was used rather than Sequence 1.

Figure 13 shows the averaged subjective estimates, based on the degraded matrix, of the probability of the hypothesis confirmed by the data. The solid curve is the objective estimate, the upper curve is for estimates made when the basic matrix preceded the degraded, and the dotted curve is for estimates made when the degraded matrix came first. All curves in Fig. 13 are based on the same sequence of messages to the subject.

Figure 14 provides the same information as Fig. 13 except that it is based on Sequence 2 rather than Sequence 1.

Figure 15 compares the objective probability estimates using the basic matrix for Sequences 1 and 2.

Figure 16 compares the objective probability estimates using the degraded matrix for Sequences 1 and 2. These last two figures make it possible to directly compare the rate of change of probabilities for the two sequences. It should be remembered that the sequences increase the probabilities for different hypotheses and are drawn with respect to the probable validity of the hypotheses which they respectively confirm.

An analysis of variance was done on the subjects' final estimates of the probability of the hypothesis that tended to be confirmed by the particular data sequence used. A separate analysis was done for scores on the basic matrix and for scores on the degraded matrix; that is, they were treated as separate scores and the order of matrix presentation was taken as an experimental treatment. The results of the analyses of variance are summed up in Tables V and VI.

TABLE V. SUMMARY OF ANALYSIS OF VARIANCE OF FINAL ESTIMATES OF PROBABILITY USING BASIC MATRIX

Source of Variation	df	MS	F	P
Columns (data sequence)	1	520.2	2.13	
Rows (matrix order)	1	1,065.8	4.37	P < .10
(cells)	3	636.9		
Rows x columns	1	924.8	3.79	P < .10
Within cells	16	243.65		
Total	19			

TABLE VI. SUMMARY OF ANALYSIS OF VARIANCE OF FINAL ESTIMATES OF PROBABILITY USING DEGRADED MATRIX

Source of Variation	df	MS	F	P
Columns (data sequence)	1	68.45	1.09	
Rows (matrix order)	1	858.05	13.67	P < .005
(cells)	3	446.85		
Rows x columns	1	414.05		
Within cells	16	62.75	6.59	P < .025
Total	19			

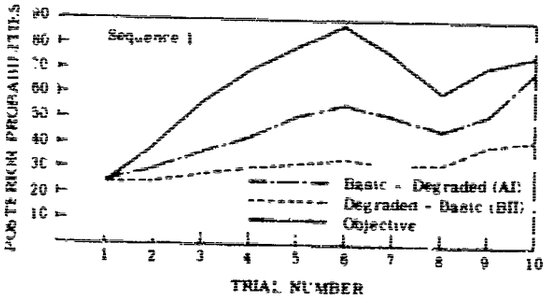


FIGURE 11. AVERAGED SUBJECTIVE ESTIMATES OF THE PROBABILITY OF THE HYPOTHESIS CONFIRMED BY DATA, USING THE BASIC MATRIX: SEQUENCE 1

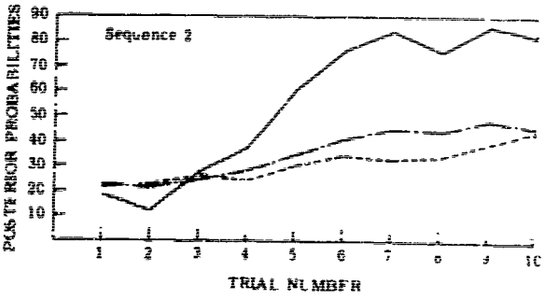


FIGURE 12. AVERAGED SUBJECTIVE ESTIMATES OF THE PROBABILITY OF THE HYPOTHESIS CONFIRMED BY DATA, USING THE BASIC MATRIX· SEQUENCE 2

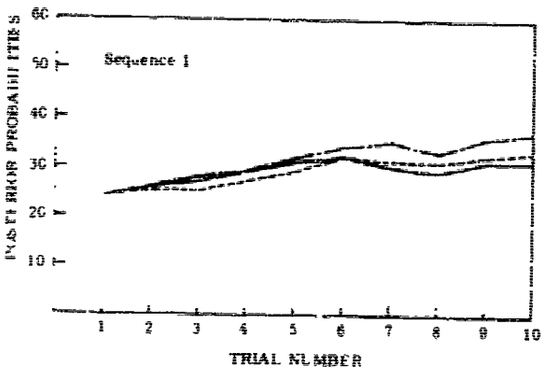


FIGURE 13 AVERAGED SUBJECTIVE ESTIMATES OF THE PROBABILITY OF THE HYPOTHESIS CONFIRMED BY DATA, USING THE DEGRADED MATRIX: SEQUENCE 1

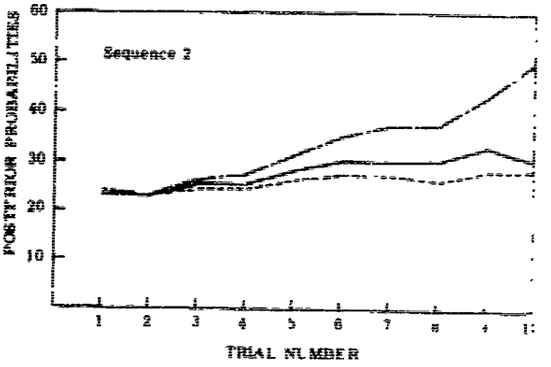


FIGURE 14. AVERAGED SUBJECTIVE ESTIMATES OF THE PROBABILITY OF THE HYPOTHESIS CONFIRMED BY DATA, USING THE DEGRADED MATRIX: SEQUENCE 2

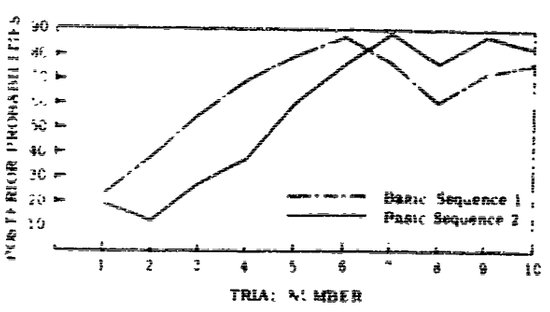


FIGURE 15 OBJECTIVE POSTERIOR PROBABILITY ESTIMATES, USING THE BASIC MATRIX

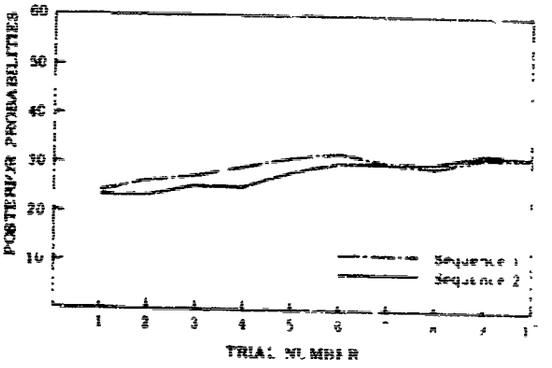


FIGURE 16. OBJECTIVE POSTERIOR PROBABILITY ESTIMATES, USING THE DEGRADED MATRIX

Separate analyses of variance on scores obtained from the basic matrix and scores obtained from the degraded matrix were done. The separate analyses preserve the effect of matrix-presentation order as an experimental treatment, and thus do not ignore an important independent variable.

As the primary question raised in the experiment was the possibility of devising a conditional probability matrix that would result in subjects' changing their probability estimates too much, the final estimation scores for the sequences run on the degraded matrix were examined by way of a t-test. Here the hypothesis tested was that the difference between the means of the final estimates and the objective value at that point differed significantly from zero; since the alternative hypothesis was that of overestimation, a one-tailed test was appropriate. The results of the t-tests are summarized in Table VII.

TABLE VII. RESULTS OF t-TESTS FOR THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN MEAN SUBJECT ESTIMATES AND OBJECTIVE ESTIMATES

		t Value	P
Sequence	B-D	3.51	$P < .005$
	D-B	.40	$P < .25$

The table clearly indicates that when the degraded matrix is used, scores show a significant overestimate for the relevant hypothesis (where overestimation is taken to be an estimate greater than the normative Bayesian probability), when the degraded matrix is presented after the basic matrix.

3.4. DISCUSSION

Figures 11 and 12 provide more evidence of underestimation of objective probabilities due to subjects' estimates changing less than is called for by Bayes's theorem. The amount of underestimation seems to be directly related to the order of matrix presentation. Both Figs. 11 and 12 show that the degree of underestimation obtained on the basic matrix when it was preceded by the degraded matrix is of a greater magnitude than that produced by the opposite presentation order. The effect of matrix order on the level of estimation is more dramatically displayed by Figs. 13 and 14, which present the scores for the degraded matrix condition. When the degraded matrix was followed by the basic, the subjects overestimated the probabilities, when the degraded matrix was presented first, subjects again tended to underestimate the probabilities. The underestimate obtained in this condition attests to the persistence

of the phenomenon; the objective probability was 0.32, but subjects managed to underestimate the relevant hypothesis and yet favor it over the others. The estimates were between 0.25 and 0.32 for seven out of the ten subjects run on this condition.

Since overestimation on the degraded matrix was only obtained when the degraded matrix was preceded by the basic it seems that the larger magnitude of estimations on the basic matrix introduces a response set that carries over into the degraded condition. This response set also seems to carry over from the degraded to the basic condition, as is indicated by the greater degree of underestimation on the basic matrix when it is preceded by the degraded.

The results of this study suggest that conservatism is found only when high-variance conditional probability displays are used. Data that has relatively low diagnostic value leads subjects to make probability estimates that are very nearly Bayesian. Under these conditions, subjects' faculties for estimating probabilities are not as bad as they would at first seem. It may well be that even for high-variance conditional probabilities subjects estimate probabilities much better than their responses indicate. This possibility gains weight from the difficulty one has conceiving a situation in which a person behaves as a pure estimator of probability, without taking other decision-making parameters into account. Conservatism may be accounted for in terms of the utilities introduced into the task of estimating probability; while the situation in this study was only a simulated war game, subjects did tend to become engrossed in the task. Their concern with the consequences of their probability estimates could, and undoubtedly did, influence those estimates to a degree. In further pursuing this line of thought, it would seem that experimental manipulation of the utilities inherent in an estimation task would answer some of the questions concerning the ability of humans to behave as "pure" probability estimators.

4
THE ESTIMATION OF CREDIBLE INTERVALS[6]

A criticism that may be leveled at many of the probabilistic information-processing experiments is that they deal with discrete hypotheses rather than continuous parameters. For example, in the experiments already described, subjects were asked to give the posterior probabilities of discrete hypotheses; they were asked to make point estimates. In the present experiment subjects were asked to estimate a continuous parameter, to give the 90% or 50% credible interval of a posterior distribution. Subjects were presented with a sequence of numbers drawn from a normal distribution with known variance but unknown means, and after each presentation of a number were required to estimate either a 90% or 50% credible interval for that mean.

It seemed that the conservatism found for discrete hypotheses might reasonably be expected in the estimation of continuous parameters also. Therefore, it was anticipated that the credible intervals given by subjects would not decrease in size with the square root of the number of observations, as they should, but would decrease more slowly.

4.1. METHOD

4.1.1. SUBJECTS. Five male summer school students at The University of Michigan volunteered to participate in the experiment. They were paid $1.25 per hour.

4.1.2. INSTRUCTIONS TO SUBJECTS. Subjects were asked to make guesses about the average or mean of a set of normally distributed numbers. They were told that they would see a sequence of numbers randomly chosen from that set and that the experimenter was interested in the degree of certainty each new number gave them about the average or mean of the set from which the sequence of numbers was drawn.

The subjects were asked to show their certainty by giving credible intervals within which they were either 50% or 90% sure that the mean should fall. They were told that as they saw more and more numbers they should become increasingly certain about the mean, and thus should be able to make their credible intervals smaller and smaller.

The subjects received instruction about the parameters of a normal distribution and its symmetry. Before seeing any numbers they were told the standard deviation of the population from which the numbers were drawn and the experimenter set an a priori credible interval within which, without seeing any numbers, they could be 50% or 90% certain the population mean would fall.

[6]This section was prepared by Marilyn F. Zivian and Ward Edwards, on the basis of data collected by Samuel M. Rubin.

They were informed that there was a perfect performance to which their performance would be compared and that he who performed best would receive an extra payment for participating in the experiment.

4.1.3. SEQUENCES. Three sequences of 64 numbers each, which we will call "original sequences," were generated by selecting numbers at random from a table of random numbers. The numbers came from a normally distributed population with mean zero and standard deviation one. From the original sequences nineteen secondary sequences were generated by multiplying each number in a sequence by one of two standard deviations (5 or 10) and adding to each number one of four mean values (0, 4, 50, or 54). The nineteen secondary sequences were labeled with letters of the alphabet from A to S and were shown in a different random order to each subject. For sequences A to P, subjects were asked to estimate 90% credible intervals, for Q to S, 50% credible intervals.

4.1.4. DISPLAY OF SEQUENCES. The sequences were displayed to the subjects on long rolls of adding machine tape which passed a window in a screen about three feet in front of the subject. A subject was shown a number, he made his estimate, and then the next number was rolled into view. Once he saw a number, it stayed in view until all 64 numbers were visible in the window.

4.1.5. PRIOR SETTINGS. When subjects were asked to give 90% credible intervals, the a priori interval set by the experimenter was that interval about the mean equal to M ± (1.645) (s.d.); for the 50% credible interval estimation, the a priori interval setting was at M ± (0.674) (s.d.).

4.1.6. RESPONSE APPARATUS. The response apparatus consisted of a wooden stand upon which were two pointers that could be moved along a calibrated scale. Different scales could be mounted on the apparatus. Each subject was asked to place the two pointers along a scale to indicate his certainty (50% or 90%) that the population mean lay within the interval he set.

4.1.7. SCALES. For each sequence subjects indicated their credible intervals on one of four scales. Each scale was calibrated in unit intervals.

Scale 1 ranged from -30 to +30. It was used conjunction with sequences of standard deviation 10 and population mean 0 or 4. Subjects estimated 90% credible intervals on this scale.

Scale 2 ranged from +20 to +80. It was used for sequences of which the population mean was either 50 or 54 and standard deviation was 10. Subjects estimated both 90% and 50% credible intervals using this scale.

Scale 3 ranged from -15 to +15. It was used for sequences of which the true mean was 0 or 4 and the standard deviation was 5. Subjects estimated 90% credible intervals on this scale

Scale 4 ranged from +35 to +65. It was used in conjunction with sequences of standard deviation 5 and mean of 50 or 54. Subjects estimated 90% credible intervals using this scale.

Table VIII summarizes the information about the sequences and scales used in the experimental design.

TABLE VIII. SCALES AND SEQUENCES

Secondary Sequence	Original Sequence	Standard Deviation	Mean	Credible Interval	Prior Setting	Scale
A	1	5	0	90%	M ± 8	3
B	1	5	4	90%	M ± 8	3
C	1	5	50	90%	M ± 8	4
D	1	5	54	90%	M ± 8	4
E	1	10	0	90%	M ± 16	1
F	1	10	4	90%	M ± 16	1
G	1	10	50	90%	M ± 16	2
H	1	10	54	90%	M ± 16	2
I	2	5	0	90%	M ± 8	3
J	2	5	4	90%	M ± 8	3
K	2	5	50	90%	M ± 8	4
L	2	5	54	90%	M ± 8	4
M	2	10	0	90%	M ± 16	1
N	2	10	4	90%	M ± 16	1
O	2	10	50	90%	M ± 16	2
P	2	10	54	90%	M ± 16	2
Q	1	10	50	50%	M ± 7	2
R	2	10	50	50%	M ± 7	2
S	3	10	50	50%	M ± 7	2

6.1.8. PROCEDURE. Subjects were run one at a time for five experimental sessions of one hour each. They saw sequences A to P and completed their 90% credible interval estimations before seeing sequences Q, R, and S and making 50% credible interval estimations. Subjects saw the sequences in the random orders given in Table VIII.

4.2. RESULTS

The widths of subjects' estimated credible intervals were analyzed by comparing them to the Bayesian interval width. Bayesian intervals were found by calculating (3.29) (s.d.)/$\sqrt{N}$ for the 90% credible intervals and (1.348) (s.d.)/$\sqrt{N}$ for the 50% credible intervals, where N = the number of the sample or trial numbers in the sequence. Plots of the comparisons showed no learning; subjects were no more Bayesian on late sequences than on early sequences. Therefore, the results were combined over all sixteen sequences for which subjects gave 90% credible intervals and over the three sequences for which subjects gave 50% credible intervals. Since plots showed large and consistent individual differences, results were not combined over subjects. Figure 17 shows the results of this analysis. Only Subject Four set intervals equal to or

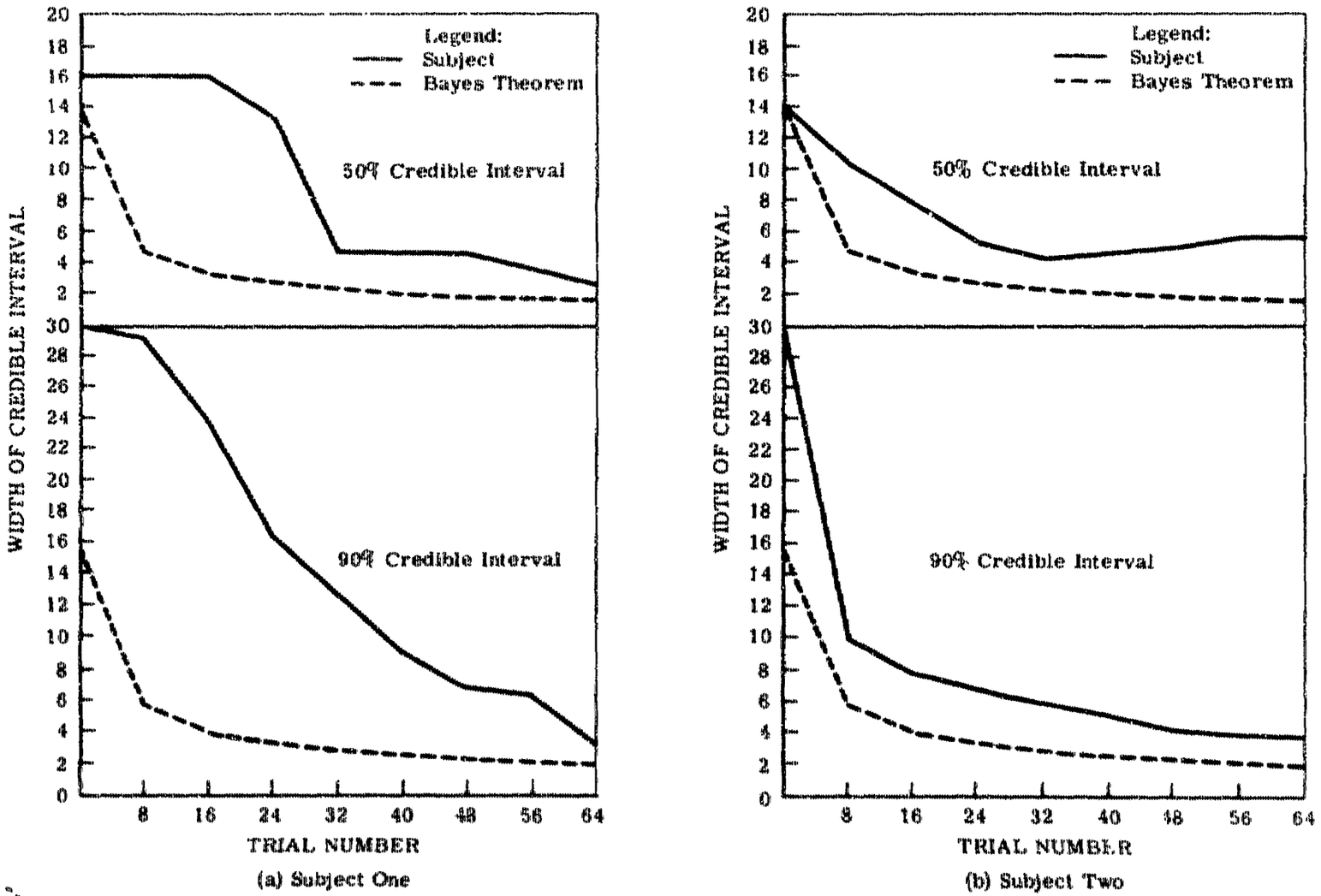
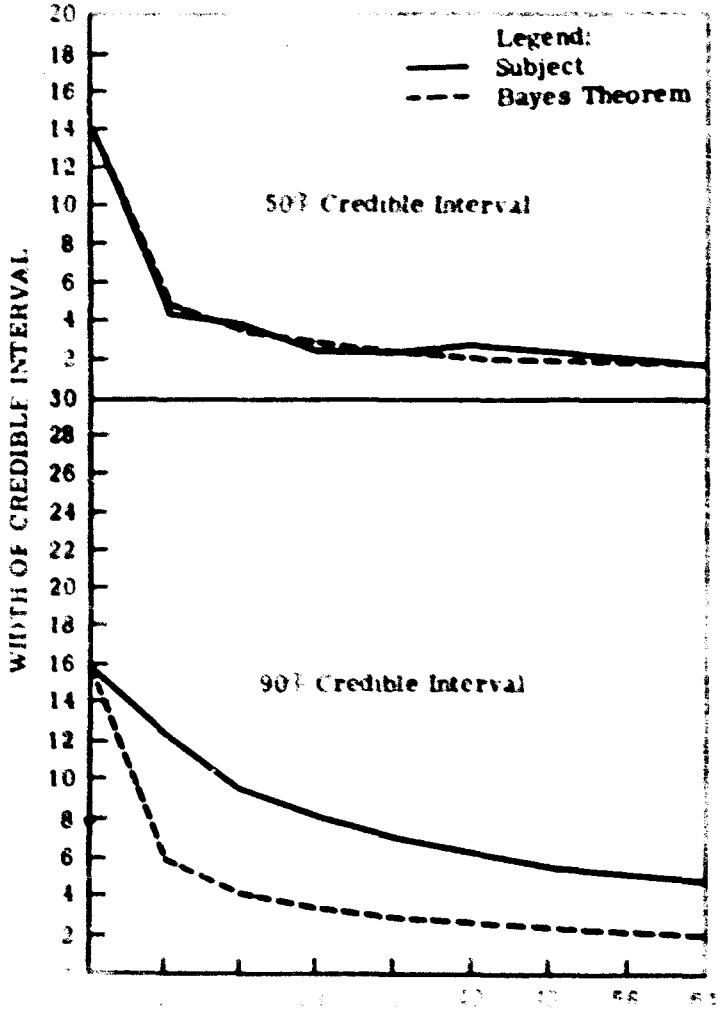
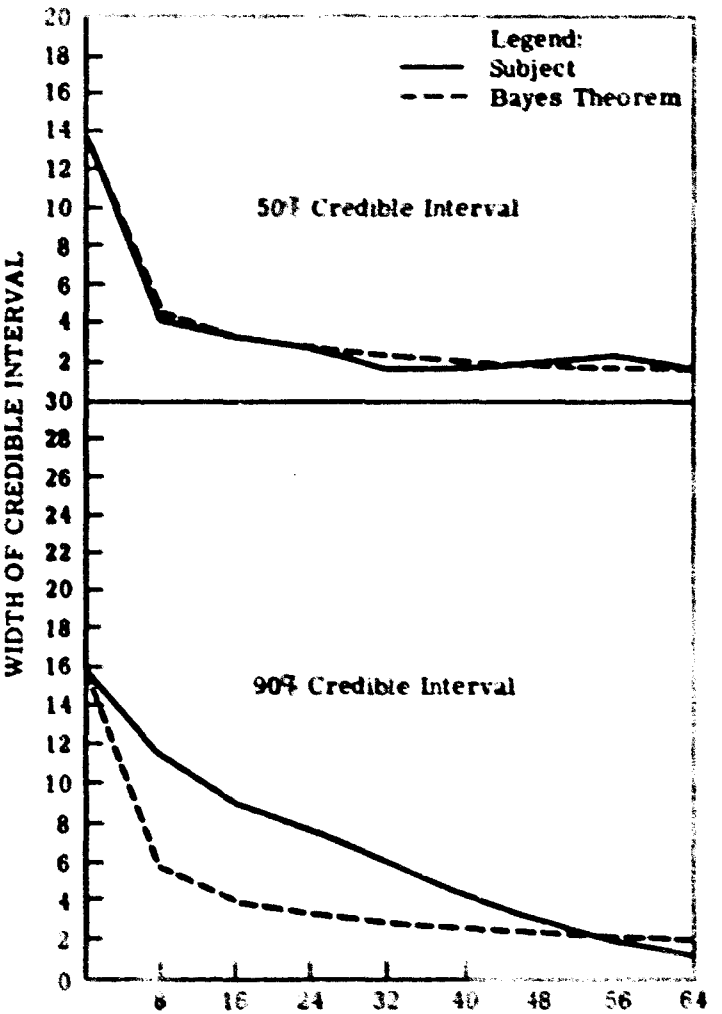


FIGURE 17. WIDTH OF CREDIBLE INTERVALS AVERAGED OVER SEQUENCES

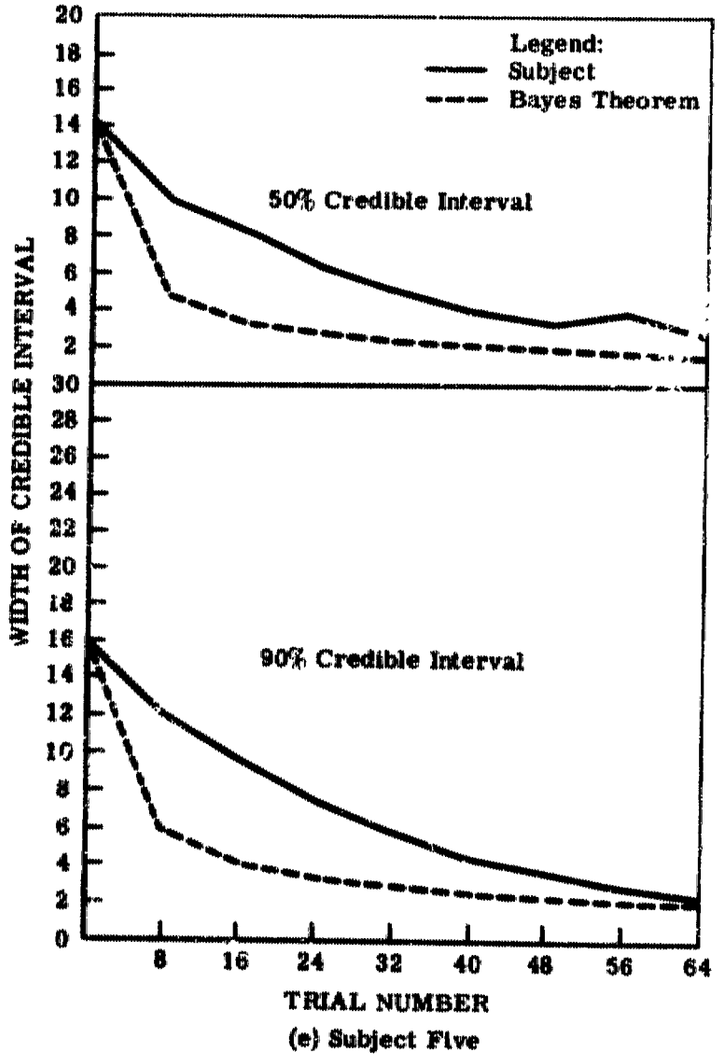


(c) Subject Three



(d) Subject Four

FIGURE 17 (continued)



(e) Subject Five

FIGURE 17 (continued)

smaller than Bayesian interval widths when giving 90% credible intervals, and only Subjects Three and Four gave intervals equal to Bayesian intervals when setting 50% credible interval widths.

The midpoints of the intervals set by subjects were used as estimates of what the subjects thought the mean of the population to be at every trial. Bayesian means were found by calculating $(\mu_0 h_0 + xh) \cdot (h_0 + h)$, where μ_0 = the prior mean, h_0 = the prior precision, x = the value of the sample, and h = the precision of the sampling process. Precisions were defined as the reciprocal of the prior-distribution variance in one case and of the sampling-process variance in the other case. The absolute deviations of a subject's means from the Bayesian means were found and summed at every trial over all 19 sequences. Figure 18 displays the summed deviations from Bayesian means at every eighth trial. A comparison of Fig. 18 with the widths of subjects' estimated credible intervals shows that there is a correlation between the subjects' ability to track the Bayesian mean and the size of the credible intervals they set.

4.3. DISCUSSION

As was expected, the subjects displayed conservatism; in seven of the ten instances examined they did not reduce their interval widths by an amount inversely proportional to the square root of N, the number of samples, but more slowly.

However, analyzing the data of the experiment pointed to problems: (1) the subjects might not have distinguished between the concepts of population mean and sample mean; (2) there is no reason why they should have believed that the numbers displayed came from a stationary process; (3) only four population means were used, two of which (0 and 50) were in the center of the scales on which subjects moved their pointers; and (4) at the beginning of each sequence, the pointers were preset by the experimenter to the theoretical size within which, without sampling, one could be 90% or 50% confident that the population mean fell, and the population mean was always at the center of this preset interval.

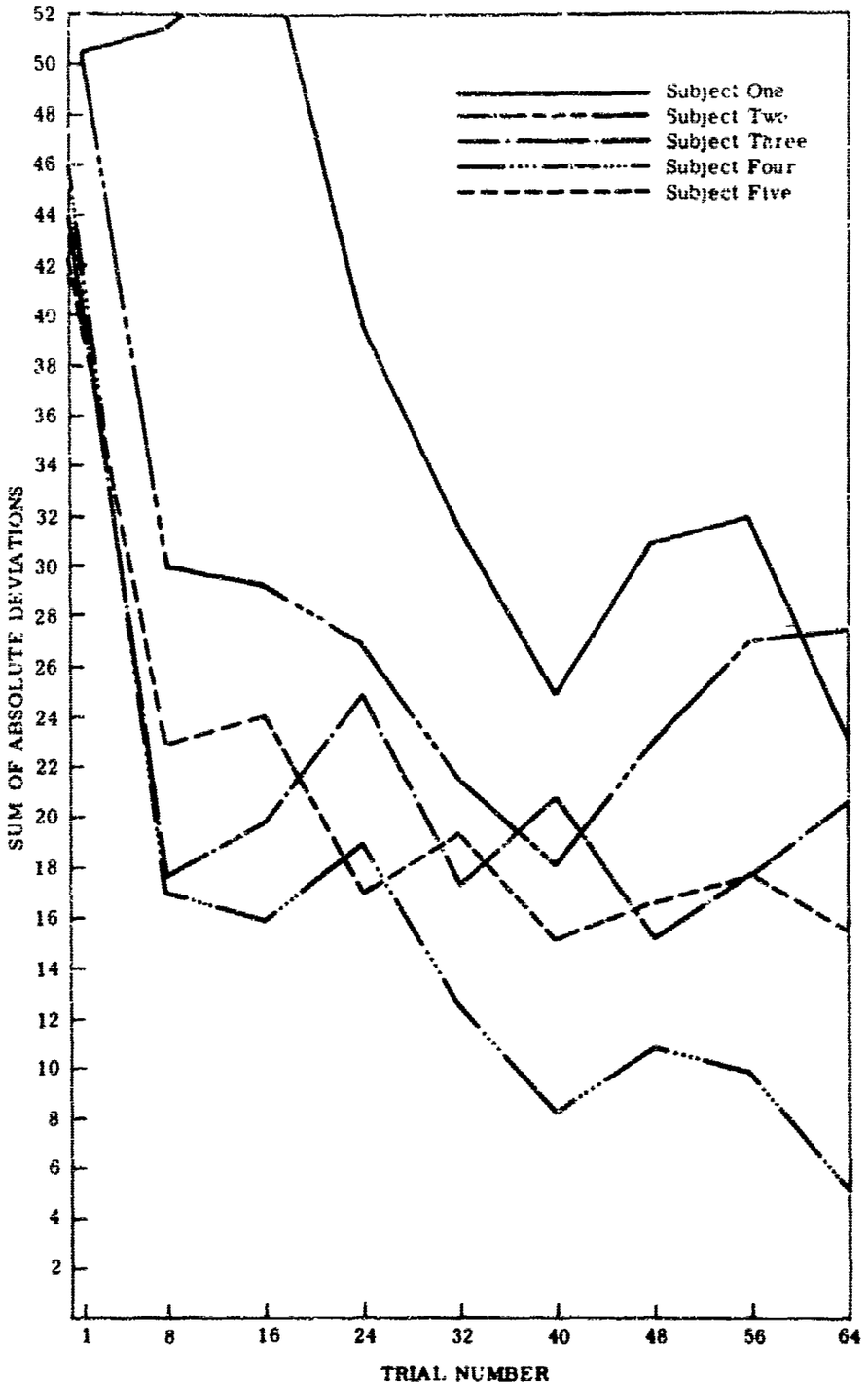


FIGURE 18. SUM OF ABSOLUTE DEVIATIONS OF SUBJECTS' MEANS FROM BAYESIAN MEANS

5

CONSERVATISM IN A VERY SIMPLE PROBABILITY-ESTIMATION TASK[7]

In the first experiment we have reported, subjects were told that the environment could be in exactly one of four possible states, referred to as hypotheses. A sequence of 15 data, generated under the truth of one hypothesis, was shown to the subjects. After seeing each datum in the sequence, the subjects estimated how probable they thought it was that each of the four hypotheses was the true one. Their estimates were compared with probabilities computed from Bayes's theorem.

The general finding of this study was that the subjects' probability estimates, while highly reliable, were considerably more conservative than those calculated from Bayes's theorem; this led to the postulation that this conservatism resulted from the intellectual difficulty of combining the diagnostic value of each individual datum in order to arrive at a diagnosis of the environment based on all the available data.

In the present study, we hypothesized that the conservatism could be reduced or even eliminated by decreasing the difficulty of the original task. In the new task, only one of two hypotheses could be true, and only two kinds of data were possible. Thus, subjects were presented with sequences of data allowing only two different observations, and only two probability estimates—one for each hypothesis—were required after subjects saw each datum. This is the simplest possible task requiring revision of opinion as new information is presented.

5.1. METHOD

5.1.1. PROCEDURE. Subjects were shown one bookbag chosen from among ten bags of which all were equally likely to be chosen. Each of the ten bags contained 100 poker chips, some red and some blue. Every bag was either a Type R bag, in which red chips predominated, or a Type B bag, in which blue chips predominated. For each type, the preponderant chips were in proportion p while the nonpreponderant chips were in proportion q. Of the ten bags, r were of Type R and b were of Type B. Subjects were told how many of the ten bags were of Type R and how many were of Type B, and they were told the exact proportions p and q.

[7] This section prepared by Lawrence D. Phillips and Ward Edwards on the basis of data collected by Richard Norman.

Subjects were told that two hypotheses about the contents of the chosen bag were possible for this experiment:

Hypothesis R: The chosen bag was Type R.

Hypothesis B: The chosen bag was Type B.

Next, subjects were asked to make intuitive estimates of the probabilities of the two hypotheses. The proportion r/10 will be called the theoretical prior probability of Hypothesis R, $P(H_R)$, and b/10 will be called the theoretical prior probability of Hypothesis B, $P(H_B)$. If the subjects' estimates differed from the theoretical prior probabilities, the experimenter explained that lack of other information made the proportions r/10 and b/10 the best estimates of the prior probabilities. This procedure ensured that all subjects started with the same prior probabilities.

Twenty chips were drawn, one at a time and with replacement, from the chosen bag. After each draw, subjects revised their previous intuitive estimates of the probability that Bag Type R had been chosen and of the probability that Bag Type B had been chosen. This process of selecting one bag at random from ten and then drawing 20 chips from the bag was repeated 24 times; thus, every subject made 20 pairs of estimates for each of 24 sequences. The correct hypothesis, the prior probabilities, and the proportion of predominant chips differed for each sequence, as shown in Table IX.

Only eight different basic sequences of red and blue chips were actually shown to subjects, as can be seen in Table IX. Sequences are apparently difficult to remember; no subject reported noticing the repetition of sequences. These sequences were

1.	FSSSS	SSSFS	FSSSS	FSSSF
2.	FSFSF	SSSSS	SSFSS	FSSSF
3.	FFSSF	FSFSS	SSSSF	SSSSS
4.	SSFSF	SSSSS	SFSSF	FFFSF
5.	SSFFF	SSSSS	SSSSS	SSSFS
6.	SFSSS	FSFSS	SSSSS	FSSSS
7.	SSFSS	SSSFS	SSSFF	FSFFS
8.	FSSSF	FSFSS	FSSSS	FSSFS

The letters S and F denote "success" and "failure", where a success is defined as the drawing of a chip with the same color as the predominant chips in the chosen bag, and a failure is the drawing of a chip of the other color. The symbol for probability of success is p, and that for probability of failure is q, and $p + q = 1$.

Sequences were presented to subjects in random order, six sequences per session. Each session lasted for about an hour. Subjects were run individually and were self-paced. Subjects were never told anything about the quality of their estimates nor were they told which hypothesis was correct for a given sequence.

TABLE IX. EXPERIMENTAL DESIGN

Sequence No.	Correct Hypothesis	$P(H_R)$	p	Basic Sequence No.
1	H_R	30	.6	1
2	H_B	30	.6	3
3	H_R	40	.6	2
4	H_B	40	.6	4
5	H_R	50	.6	3
6	H_B	50	.6	1
7	H_R	50	.6	4
8	H_B	50	.6	2
9	H_R	60	.6	2
10	H_B	60	.6	4
11	H_R	70	.6	1
12	H_B	70	.6	3
13	H_R	30	.7	6
14	H_B	30	.7	8
15	H_R	40	.7	5
16	H_B	40	.7	7
17	H_R	50	.7	8
18	H_B	50	.7	6
19	H_R	50	.7	5
20	H_B	50	.7	7
21	H_R	60	.7	7
22	H_B	60	.7	5
23	H_R	70	.7	6
24	H_B	70	.7	8

5.1.2. SUBJECTS. Five males, undergraduates of The University of Michigan, served as subjects. They were paid $1.25 per hour.

5.2. RESULTS

Theoretical probabilities for each sequence can be calculated from Bayes's theorem:

$$P(H_R \cdot D) = k\ P(D \cdot H_R)\ P(H_R) \tag{1}$$

$$P(H_B \cdot D) = k\ P(D \cdot H_B)\ P(H_B) \tag{2}$$

$P(H_R)$ and $P(H_B)$ represent the prior probabilities of the correct hypothesis; $P(H_R \cdot D)$ and $P(H_B \cdot D)$, the posterior probabilities, or the probabilities of the hypotheses after observing the datum D; and $P(D \cdot H_R)$ and $P(D \cdot H_B)$, the likelihoods of the datum or the conditional probabilities of the datum given the truth of the particular hypothesis. A normalizing constant k ensures that

$$P(H_R \cdot D) + P(H_B \cdot D) = 1$$

A form of Bayes's theorem more convenient for analyzing the data can be obtained by dividing Eq. 1 by Eq. 2 whenever H_R is the correct hypothesis, and Eq. 2 by Eq. 1 whenever H_B is the correct hypothesis. This gives,

$$\Omega_1 = L\Omega_0 \tag{3}$$

where Ω_1 represents the posterior odds in favor of the correct hypothesis; Ω_0, the prior odds in favor of the correct hypothesis; and L, the likelihood ratio of the data.

Since each draw of a chip is generated by a binomial process, with probability of success equal to p, the probability of getting s successes in n draws is proportional to $p^s q^{n-s}$. Thus, the likelihood ratio of the datum is

$$L = \frac{p^s q^{n-s}}{q^s p^{n-s}} = \frac{p^{2s-n}}{q^{2s-n}} = \left(\frac{p}{q}\right)^{2s-n} \tag{4}$$

Of course, $2s - n = s - (n - s) = s - f$ is the difference between the number of successes and failures, so Eq. 4 can be written

$$L = \left(\frac{p}{q}\right)^{s-f} \tag{5}$$

Rewriting Eq. 5 in log form gives

$$\log L = (s - f) \log \frac{p}{q} \tag{6}$$

This form is convenient because, for given values of p and q, log L varies linearly with s - f. Figure 19 shows a plot of $\log_{10} L$ as a function of s - f. Two plots are shown, one for each value of p used in this experiment.

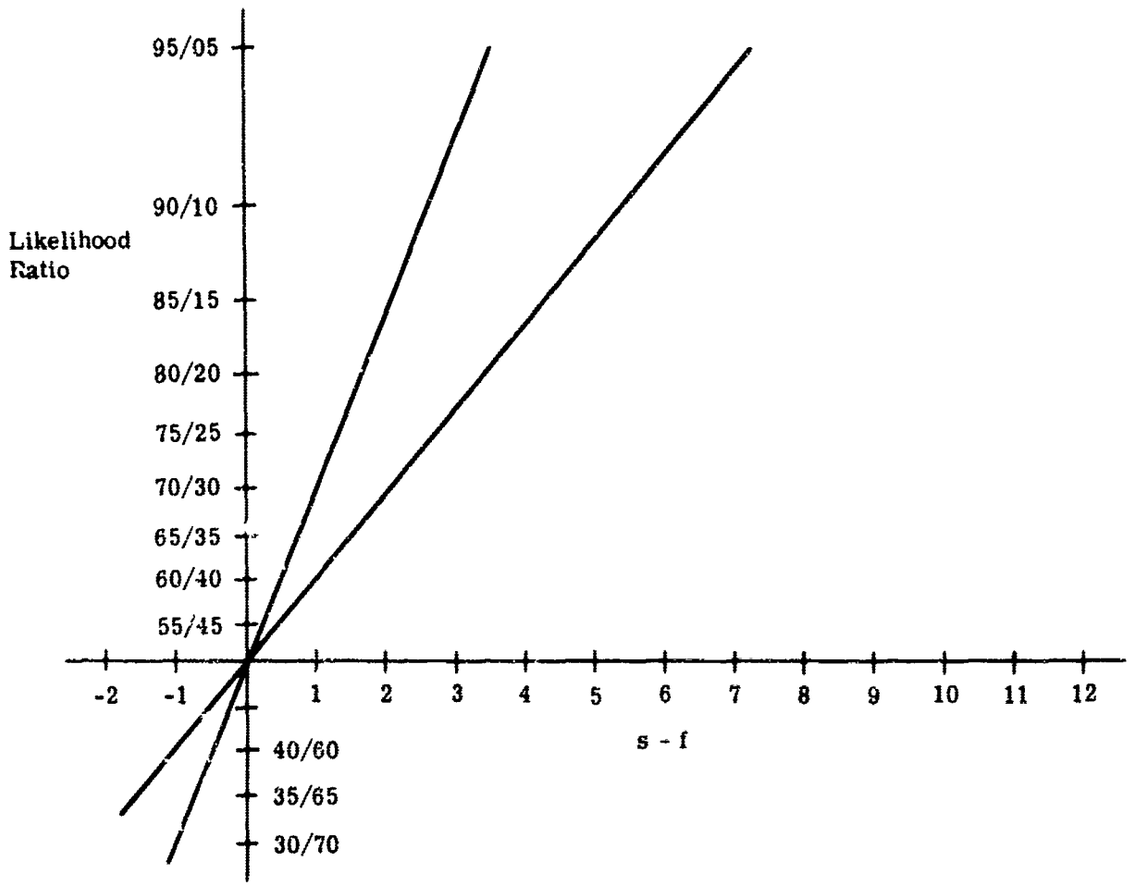


FIGURE 19. THEORETICAL LIKELIHOOD RATIOS, FOR 70-30 AND 60-40 BOOKBAGS, AS A FUNCTION OF THE DIFFERENCE BETWEEN THE NUMBER OF SUCCESSES AND THE NUMBER OF FAILURES

The log likelihood ratios computed from Eq. 6 and shown in Fig. 19 are theoretical values. Log likelihood ratios inferred from subjects' estimates can also be computed and compared to the theoretical values. First, subjects' estimates were converted to posterior odds. Then, since the prior probabilities were given, inferred likelihood ratios can be calculated from this logarithmic form of Eq. 3:

$$\log L = \log \Omega_1 - \log \Omega_0 \qquad (7)$$

Plotting subjects' likelihood ratios as a function of Bayesian likelihood ratios allows actual performance to be compared with theoretical performance. This has been done in Fig. 20 for Subject One, for the data obtained in sequences with p = .7. Plots for data obtained when p = .6 gave nearly identical results, so are not shown here. The scatterplots of all subjects except Subject Four were similar to those of Subject One; Subject Four's show greater scatter.

Another way to summarize these data is to determine what bookbag compositions would be necessary for Bayes's theorem to give probabilities identical to those estimated by the subjects. This has been done graphically, and the results are given in Table X.

TABLE X. RANGE OF p VALUES THAT WILL YIELD BAYESIAN PERFORMANCE IDENTICAL TO SUBJECTS' ESTIMATES

Subject	True Value of p	
	.7	.6
1	.51-.55	.50-.55
2	.50-.54	.50-.56
3	.52-.56	.51-.59
4	.50-.60	.50-.69
5	.50-.53	.50-.54

For example, the data generated by Subject One when he saw a 70-30 bookbag could have been generated by Bayes's theorem using values of p which ranged from .51 to .55.

5.3. DISCUSSION

Despite the simplicity of this task, subjects' estimates were still conservative, compared to probabilities computed from Bayes's theorem. Apparently, the conservatism found in Experiment One is not entirely caused by the complexity of that task.

Table IX indicates that the amount of conservatism is very little affected by the two values of p in this experiment. Possibly this is caused by presenting sequences in random order. If all the .6 sequences had been presented together, and all the .7 sequences together, perhaps the inferred likelihood ratios would have differed more.

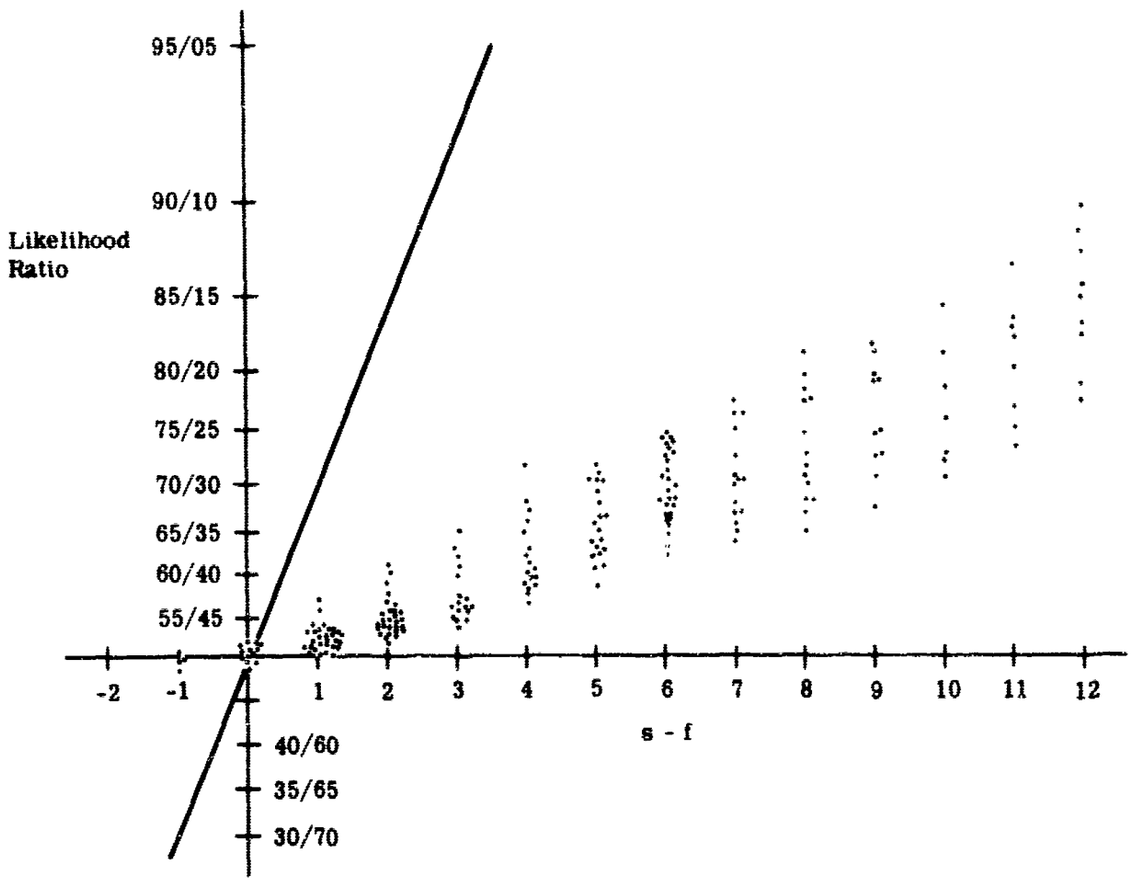


FIGURE 20. SUBJECT ONE'S ESTIMATES, FOR 70-30 BOOKBAGS, EXPRESSED IN LOG LIKELIHOOD RATIOS AS A FUNCTION OF THE DIFFERENCE BETWEEN THE NUMBER OF SUCCESSES AND THE NUMBER OF FAILURES

Finally, four of the five subjects show considerable consistency, as is indicated by the low degree of scatter in their scatterplots. Behavior in this simple task can best be described as reliable and consistent, but very conservative when compared to Bayes's theorem.

A very simple model gives a good fit to these data. It supposes that the subject raises the likelihood ratio to a power less than one before performing the arithmetic of Eq. 3; it is equivalent to saying that he behaves as though the bookbags are nearer 50-50 than they are. While this model is far too crude to be plausible, it fits these data as well as their scatter permits.

6
RESPONSE MODES AND PROBABILITY ESTIMATION[6]

Previous research (see preceding sections) has repeatedly demonstrated that subjects exhibit suboptimal behavior when processing probabilistic information, with Bayes's theorem providing the standard.

For the first experiment (Section 2) a pseudomilitary game was presented to subjects who viewed the progressive accumulation of impact points on a display that resembled a radar display (PPI), and, on the basis of these data, made posterior probability estimates about the truth of four hypotheses. The subjects consistently underestimated high probabilities and overestimated low probabilities; they were unable to extract from the information all the certainty about the truth of the hypotheses that was justifiable by Bayes's theorem.

Section 5 reports a much simpler task involving Bayesian inference. Despite the simplicity of the task, subjects were also unwilling to commit themselves to extreme probability estimates. Tasks for which posterior Bayesian probabilities were greater than 0.999 elicited from subjects estimates between 0.80 and 0.90.

This conservatism seems sufficiently certain to permit investigation into the effects on it of other variables. L. D. Phillips (unpublished) employed the same bookbag and poker chip problem but explored the effect of making payoffs to the subjects contingent upon the accuracy of their posterior probability estimates. Four groups were run, a control with no payoff, and three payoff groups in which the payoffs had either a logarithmic, quadratic, or linear relationship to the probability estimates. All subjects were more conservative than Bayes's theorem; low probabilities were overestimated and high ones were underestimated. The logarithmic and linear payoff groups were more accurate in their estimates than the control group. For some reason, however, the performance of the quadratic payoff group fell below that of the control group.

The major purpose of the present study is to investigate the relative effects on performance of various probability-estimation response modes.

6.1. METHOD

6.1.1. SUBJECTS. The subjects were 15 male students of The University of Michigan randomly assigned to one of the three experimental groups; PR, VO, and ODL. Those in Group PR made their estimates by distributing 100 washers over two pegs, which forced them to normalize their probability estimates. Subjects in Group VO reported their estimates in verbal odds

[6] This section was prepared by Mary Ann Price Swain and Ward Edwards.

in favor of the most likely bag. And subjects in Group ODL made their estimates by setting a pointer along a scale on which odds were displayed in logarithmic intervals:

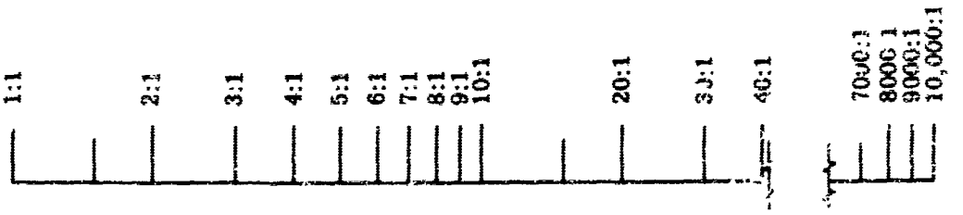


FIGURE 21. LOGARITHMIC SCALE FOR SUBJECTS' REGISTERING OF PROBABILITY ESTIMATES

6.1.2. PROCEDURE. This experiment used the bookbag and poker chip paradigm explained above. Subjects were run one at a time, and each was run in two different experimental sessions. The first session utilized 70-30 bookbags and the second session 60-40 bookbags. All bags had a prior probability of 0.5. At each session the subject was shown six different 20-chip sequences. Sequences were generated randomly and checked by the experimenter for their "representativeness." Retained sequences always favored the correct point hypothesis over the uniform hypothesis (i.e., that all compositions are equally likely); this requirement is satisfied if $(n + 1)\binom{n}{s}p^s(1 - p)^{n-s} \geq 1$, where p represents the probability of obtaining a chip of the preponderant color from the chosen bookbag; n, the total number of chips drawn; and s the number of those chips drawn that are of the color predominant in the bag. Retained sequences also satisfied the Wald-Wolfowitz test for the expected number of runs (alternation of colors) in a given sequence of s preponderant elements and n-s nonpreponderant elements.

Sequences were drawn and recorded ahead of time. During the experimental session, the experimenter presented the subject with the chips as if he were actually drawing them from a bookbag. Each subject saw the same sequences, although not in the same order. They were required to make an estimate after each draw of the sample; they were never told which was the correct hypothesis, nor were they given any feedback about the accuracy of their estimates.

6.2. RESULTS

(For reasons that will be given later, the 60-40 data failed to yield any consistent results. Therefore, the analyses to be presented here pertain only to estimates made in the 70-30 problem.)

The logarithmic odds-likelihood ratio form of Bayes's theorem is convenient for data analysis since it makes optimal performance appear linear. (This statement is fully explained under "Results" in Section 5.) Remember that this form is:

$$\log L = \log \Omega_1 - \log \Omega_0$$

where L represents the likelihood of the datum; Ω_0, the odds before observing that datum; and Ω_1, the odds after observing the datum. If we assume that the subjects have based their estimates on the values of the variable s - f, then we can compute an inferred liklihood ratio for each subject by translating each posterior estimate into its logarithm and subtracting the log of the prior odds. Figures 22-25 are typical scatterplots of subjects' inferred log-likelihood ratio. The broken line is the best-fitting regression line that passes through the origin. For all subjects, the regression lines deviate markedly from the line representing perfectly Bayesian performance. Table XI summarizes both group and individual performances. In the table, m is the

TABLE XI. SLOPE CONSTANTS, CORRELATION COEFFICIENTS, AND k VALUES FOR EACH SUBJECT AND GROUP

Group		m	r	k
PR		.081	.668	.221
	1	.062	.829	.169
	2	.094	.417	.254
Subject	3	.116	.927	.314
	4	053	.836	.145
	5	.084	.799	.228
VO		.114	.665	.310
	1	.083	.823	.225
	2	.076	.832	.207
Subject	3	.222	.573	.603
	4	.216	.847	.587
	5	.117	.945	.318
ODL		.127	.599	.345
	1	.099	.796	.268
	2	.113	.958	.307
Subject	3	.064	.677	.173
	4	.278	.842	.756
	5	.281	.976	.764

slope of the regression line, r is the measure of correlation between the s - f value and the inferred log-likelihood ratio, and k is the constant by which one multiplies the slope of the Bayes's theoretical line (log p/q) to obtain the subject's slope (m).

Table XI shows that response modes do affect performance. The odds groups are both superior to the probability estimation group. Furthermore, the ODL group is slightly superior to the VO group.

Another way to analyze these data is to calculate the percentage of improvement in performance shown by the two odds groups over the probability estimation group. Figure 26 illustrates that by the third draw the VO subjects were 43 percent more accurate than the PR subjects and the ODL subjects were 60 percent more accurate As evidence accumulates all subjects

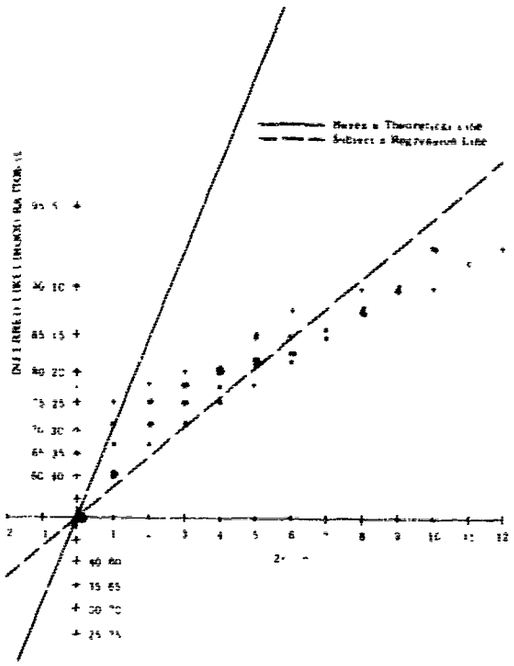


FIGURE 22. INFERRED LIKELIHOOD RATIOS FOR VO SUBJECT FIVE

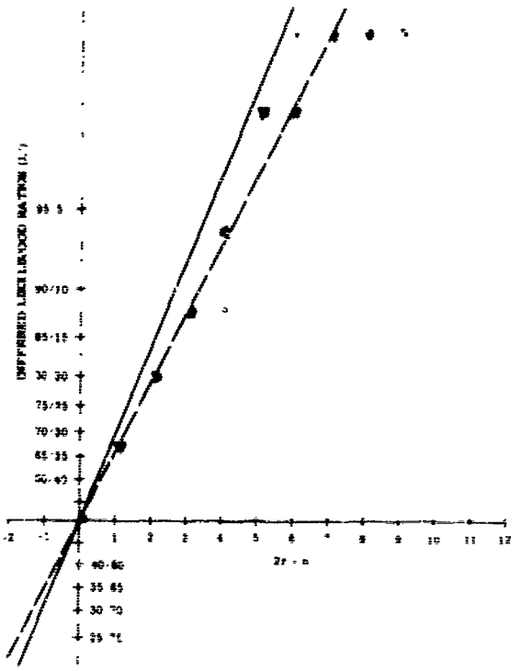


FIGURE 23. INFERRED LIKELIHOOD RATIOS FOR ODL SUBJECT FIVE

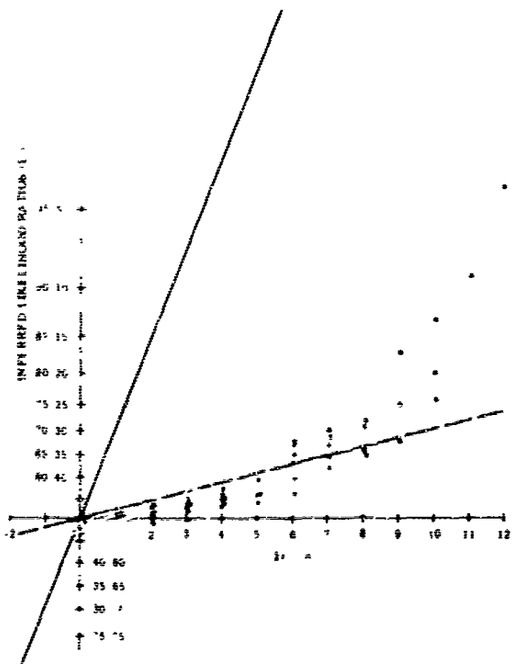


FIGURE 24. INFERRED LIKELIHOOD RATIOS FOR PR SUBJECT TWO

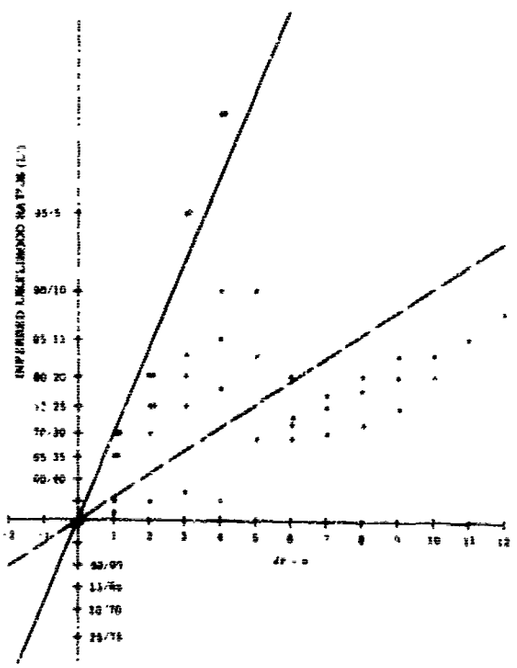


FIGURE 25. INFERRED LIKELIHOOD RATIOS FOR PR SUBJECT THREE

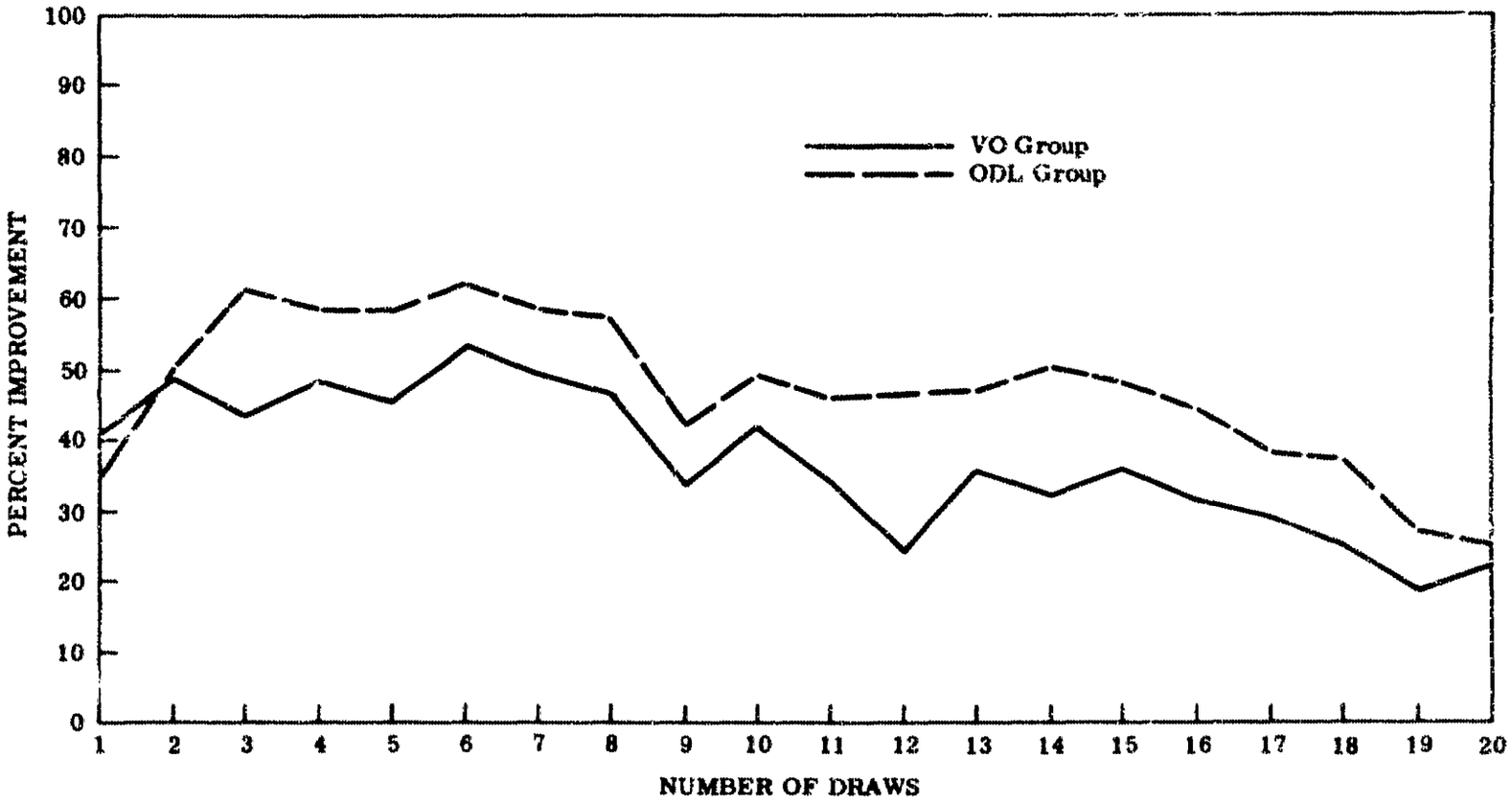


FIGURE 26. PERCENT OF IMPROVEMENT SHOWN BY VO AND ODL GROUPS OVER PR GROUP IN ACCURACY OF ESTIMATION

should increase their certainty about the truth of a hypothesis; consequently, by the 20th draw the differential performance of the groups was reduced. At this point, the VO and ODL groups were only 22 and 24 percent more accurate in their estimates.

6.3. DISCUSSION

This study reconfirms the finding that subjects are conservative in situations involving inference from fallible information, and are unable to extract from information all the justifiable certainty about the truth of a hypothesis. The roughly linear scatterplots in Figs. 22-24 are characteristic of the majority of the subjects. Occasionally (Fig. 25) a subject will exhibit great variability in his estimates. In this case, the subject told the experimenter that he had changed his strategy in the middle of the session. A simple model to describe such a subject's behavior is

$$\log L' = k \log L$$

where L' represents the subject's inferred likelihood ratio; the values of k are shown in Table XI.

Subjects are Bayesian information processors, but they raise every likelihood ratio to a power less than one. Another way of describing this model is that subjects behave as though they do not believe the experimenter's statement about the composition of the bookbags. PR subjects behaved as if they thought the bags were of a 55-45 composition; VO, 56-44; and ODL, 57-43. In short, subjects degrade the environment in a consistent way.

If subjects are hesitant to commit themselves to extreme estimates, then one would expect the performance of those who estimate odds to be superior to that of those who estimate probabilities, because probabilities have an upper limit of 1.00. Thus, as PR subjects increase their estimates they also reduce the upper range of responses remaining to them. Odds do not have this upper limit. Therefore, it is easier for the VO and ODL groups to make larger estimates since they always have an unlimited range of estimates still available. Moreover, the visual logarithmic display of odds further facilitates making large estimates.

Phillips found that paying subjects for accuracy tended to enhance their performance. The results of this experiment suggest that subjects should estimate posterior odds rather than posterior probabilities in an information-processing task. It would be convenient if the positive effects of payoffs and odds combined additively to influence total performance. That, however, is an experimental question to be explored.

The data from the 60-40 sequences were not analyzed for the following reasons: three out of five subjects in the VO group and four out of five in the ODL group gave as their odds estimates the ratio between the number of red chips and the number of blue chips presented to them. Secondly, one subject in PR, two subjects in VO, and one subject in ODL told the experimenter that they felt confused in the 60-40 case since they were still thinking of 70-30 bookbags. The data, consequently, are ambiguous and difficult to interpret.

This study should be repeated, employing more than one Bernoulli probability for the bookbags. Since subjects behave as if the composition of the 70-30 bookbags were in the vicinity of 55-45, it would be of interest to see if they are more nearly Bayesian information processors when the actual bookbag composition is 55-45. A more extreme p value should be chosen (0.85 or 0.9) in order to see if the differential performance of the two odds groups is maintained. Asymmetric bookbags would further test various response modes. In any case, careful controls should be exercised to insure that subjects do not confuse an odds estimate with the sample ratio of red chips to blue chips.

Appendix A
INSTRUCTIONS TO SUBJECTS

Suppose you are in the Air Force and stationed at one of their radar detection stations in Greenland. These stations have large, powerful radars that detect many types of aerial activity — ICBM's, rockets, planes, clouds — sometimes even birds. All of these things may show up on the display — the radar scope. Unfortunately by the time they are displayed they may look alike — little spots of light on a dark background. Obviously, you have a problem if you happen to have the job of sitting at one of these scopes and trying to figure out what are enemy ICBM's and what are birds. Fortunately, the problem isn't hopeless. For instance, in the example just given, the ICBM's versus the birds, ICBM spots would obviously move faster than birds.

You're not here so we can train you to be a good radar operator in case you should ever find yourself in Greenland; however, the series of experiments in which you are about to participate does concern the problem of evaluation.

Although the information presented to you will be in simplified form, the basic elements of the problem will be very similar to an actual situation. You will play the part of an evaluator: it will be your job to decide among four possible types of airborne activity (POINT TO CONSOLE): enemy, friendly, meteor, or spoof. Enemy activity may be of any sort, an ICBM or rocket, for example. For the purpose of this experiment the specific type of enemy threat is not important. Friendly activity may also be of any sort. Meteors are self-explanatory. A spoof is a diversionary or probing activity by the enemy, like the cowboy hero who throws his hat in the air to see what the bad guys will do about it.

You are seated at the output display of a complex detection system. This detection system covers a large, circular area that will be subdivided, for this problem, into sectors. This area will be displayed here. (TURN ON SECTOR DISPLAY — A SLIDE WITH NO IMPACT POINTS).

Aerial activity is detected by means of a powerful radar system. radar information on detected targets is fed to a computer that determines the courses and speeds of the targets and the paths they are following. For this experiment, it will be assumed that the courses and speeds of the targets do not change once detection is made. Once the courses and speeds and the paths of the targets are determined, the computer determines where the targets will land. These points of impact will be displayed on the console within one of the sectors of this land

area. Since we obviously don't really have a radar here, a 35-mm slide projector projects this display from the back of the console. (DISPLAY SLIDE WITH SEVERAL IMPACT POINTS). To simplify this experiment, we have not put any dimensions on this circle of land area; just consider it as a land mass on which points of impact are displayed. Remember, computed impact points are being displayed here, not the radar targets themselves.

For each experiment you will be shown fifteen slides. In some experiments the number of impact points will increase with each successive slide. In others, the number of impact points will change erratically with each successive presentation. For both of these types of experiments, the impact points on one slide are all of the same type of activity. Thus, regardless of whether there are three or thirteen impact points displayed here by one slide, they all represent the same type of activity, that is, they are all enemy, or all friendly, or all meteors, or all spoof, not a combination. However, the two types of experiments differ in this respect: in the series where the number of impact points successively increases, the activity represented on one slide is the same as for the previous slide. For the erratic series, each slide of the fifteen may represent activity different from that on the previous slide.

To summarize then, there are two types of experiments in which you will be involved. In one type you will first be shown one computed impact point (SHOW), then one more (SHOW), then another (SHOW), and another (SHOW), and so on until fifteen presentations (SHOW) have been made. The impact points on any one of these slides represent all the same activity and the activity represented by each slide is the same as that on the previous slide. Thus each and all of these slides just shown may have represented friendly activity. In the other type of experiment, first you may be shown, for example, three impact points (SHOW) representing a single kind of activity. The next slide may have eleven impact points (SHOW), again all of the same activity. However, the activity represented by this slide may be different from that of the previous slide. Thus, the previous slide of three impact points may have represented enemy activity, while this one represents meteors. Fifteen presentations will be made for this type of experiment, also.

Before you begin each experiment, you will be told whether the displayed impact points represent changing activity or the same activity. Incidentally, slides in both experiments will be of the type you see here, that is, white impact points on a black background. Are there any questions on what is to be displayed?

It will be your problem to decide which of the four types of activity is being displayed by the computed impact points. To help you in this evaluation, five pieces of information will be given to you.

First, we will assume that through advance intelligence you have some estimation of how likely an enemy attack may be. We will limit the experiment to three possible estimations:

1-in-10 chance of enemy attack, 1-in-4 chance, or 2-in-3 chance. That is, you will be told that there is either a 10% likelihood of enemy attack, or a 25% likelihood, or a 67% likelihood. (SHOW BASE RATES, INSERT 25%). The second piece of information will give you an idea of where an enemy impact point is likely to be. (INSERT ENEMY DISPLAY). This display shows, in percentages and in pie diagrams, what probability there is that an enemy missile will land in any one of the sectors. Here, the probability is highest in this 25% sector and lowest in this 2% sector. In other words, if the impact points are those of an enemy, they are more likely to show up in the sectors with the higher numbers, or with the bigger pie slices. The third, fourth, and fifth pieces of information are similar displays for friendly, meteor and spoof activity. (INSERT THEM WHILE EXPLAINING). You will notice that there is a rough pattern to each of these possible types of activity. (POINT TO PATTERNS). Enemy attack generally would come from this direction; friendly activity would more likely be concentrated in this area; meteors would probably be found here; spoof activity would tend to be in this area.

One important point should be mentioned now. Although the pie diagrams are shown near the center of each sector, the percentage each represents applies evenly to the whole sector. (POINT TO 5% SECTOR). In other words, this 5% value applies evenly to this whole sector. Thus the dividing line between sectors represents a sharp change in likelihood; there is no gradual shading from one likelihood to another. Remember, then, each sector is of constant likelihood.

In summary, you will evaluate the type of activity represented by a set of impact points. Five pieces of information will be available to use as you desire: the likelihood of enemy attack; the likelihood that, if friendly activity is being observed, the computed impact points would appear in certain sectors; and similarly for meteors and spoofs. You will make an evaluation after the display of each slide. Thus, for one experiment, you will make fifteen evaluations. (CHANGE TO BLANK SLIDE).

Your decisions will be made with the levers on the console. The numbers to the left of each lever indicate your estimates of the likelihood that the impact points represent the corresponding type of activity. The lower end, near zero, represents very low likelihood, the upper end, near one, represents very high likelihood. If you set the ENEMY lever to .6 (SET LEVER) this means you estimate that there is a 60% probability, or likelihood, that the impact points shown here represent enemy activity. (RETURN LEVER TO ZERO). After the first slide has been displayed, make your evaluation of the type of target represented by the impact point, or points. Indicate your probability estimates by moving the levers to the appropriate levels.

For instance, if you moved the levers to .6, .1, .25, and .05 (MOVE LEVERS ACCORDINGLY), this would indicate that you believe that there is a 60% probability that the impact points

represent enemy activity, 10% probability they are friendly, 25% probability they are meteor, and 5% probability they are spoof.

Let's look at what I've just said from a little different point of view. Before you start an experiment, the best estimate we have of the probability of enemy attack is this advance intelligence statement. All we know is that there is a 25% probability that enemy missiles will appear. Additionally, this display (POINT TO ENEMY DISPLAY) tells us that if the enemy attacks, his missiles are likely to fall in this way, and similarly for the other three types of activity.

So you see, we're dealing with three types of probability estimates. One is given before the experiment starts: it is a statement of what to expect. Another, shown on these cards, (POINT TO P(D|H) DISPLAYS), says "if it happens, the impact points are likely to fall like this, (POINT TO ENEMY DISPLAY), and if it doesn't happen, the impact points are likely to fall like this (POINT TO ANY OTHER DISPLAY)." And the third is your estimate of what is actually happening.

Now, are there any questions so far?

The console is operated by this white button. When the green light is lighted, pushing the button will cause the display here to be revealed. I have already done this. Then, you make your evaluation and set the levers. When you are finished push the button — go ahead, try it. The red light comes on, indicating that the lever settings are being recorded on a special recorder behind the console. You mustn't move the levers while the red light is on. When the lever settings have been recorded, the yellow light comes on. This is a signal for you to reset the levers to zero. Try it. When they are all reset, the green light comes on. If the yellow light stays on, check the position of all four of these levers again, as well as these extra two. The zero point is quite sensitive, and sometimes the levers are jarred off this position.

As soon as the green light comes on, you can push the button again to reveal the new impact-point display. Now try the sequence for yourself. Make a meaningless evaluation. (WHEN GREEN LIGHT COMES ON, STOP SUBJECT). Notice that if you accidentally move one of the levers off the zero position before a new slide comes on, the green light will blink. Resetting the offending lever will cure the situation.

Finally, you don't have to count the number of slides in the experiment. The screen will show all black when you are finished. (TURN ON BLANK SLIDE). When this happens, let me know. I'll be in the next room. There is no time limit on any of these experiments, but you should, after running through a few sets, complete a set of fifteen slides in less than fifteen minutes.

Now, are there any questions on any aspect of the experiment? For this first set, I'll stay here with you to answer any other questions which may come up as you work the console.

Appendix B
PUBLICATIONS

This appendix summarizes publications already produced under Contract AF 19(604)-7393. It confines itself to publications that have appeared in journals or as technical documentary reports, or that have been accepted and are scheduled to appear in some such form, plus one Ph.D. thesis. In the course of Contract AF 19(604)-7393, approximately 25 speeches reporting contract research were given at various formal and informal meetings. Although the more formal speeches qualify as publications also, no attempt is made in this report to list them. The technical content of every speech parallels some written technical documentary report.

This list of publications is an important complement of the present report, since no attempt has been made in the final report itself to repeat already-published ideas. The body of the final report is devoted only to the presentation of materials not yet published.

1. Edwards, W., "Dynamic Decision Theory and Probabilistic Information Processing," Human Factors, 1962, 4, 59-73.

   This paper is essentially a program review as of 1961. The development of a dynamic decision theory will be central to the impending rapid expansion of research on human decision processes. In a taxonomy of six kinds of decision problems, five require a dynamic theory in which the decision maker is assumed to make a sequence of decisions, basing decision n + 1 on what he learned from decision n and its consequences. Research in progress on information seeking, intuitive statistics, sequential prediction, and Bayesian information processing is reviewed to illustrate the kind of work needed. The relevance of mathematical developments in dynamic programming and Bayesian statistics to dynamic decision theory is examined. A man-computer system for probabilistic processing of fallible military information is discussed in some detail as an application of these ideas and as a setting and motivator for future research on human information processing and decision making.

2. Edwards, W., "Men and Computers," in R. M. Gagne (Ed.), Psychological Principles in Systems Development, Holt, Rinehart and Winston, 1962, 75-113.

   This expository chapter explains what a computer is and how it works, discusses programming and programming languages, reviews the technology of the man-computer interface, and illustrates real-time, on-line use of computers in a hypothetical information-processing system.

3. Hays, W. L., "On Lattice Models for Psychological Scaling," Psychometrika, in press.

4. Edwards, W., Probabilistic Information Processing in Command and Control Systems, ESD-TDR-62-345, IST Report No. 3780-12-T, University of Michigan, Institute of Science and Technology, Ann Arbor, 1963, 34 pp.

This is the basic document about PIP. It discusses the diagnostic function in command and control systems, and presents Bayes's theorem, examines its role in the design of command and control systems that probabilistically process fallible information. After summarizing existing relevant experimentation, the report points out major unsolved technical problems and outlines a program of research for solving some of them.

5. Edwards, W., Lindman, H., and Savage, L. J., "Bayesian Statistical Inference for Psychological Research," Psychol. Rev., 1963, 70, 193-242.

Bayesian statistics, a currently controversial viewpoint concerning statistical inference, is based on a definition of probability as a particular measure of the opinions of ideally consistent people. Statistical inference is modification of these opinions in the light of evidence, and Bayes's theorem specifies how such modifications should be made. The tools of Bayesian statistics include the theory of specific distributions and the principle of stable estimation, which specifies when actual prior opinions may be satisfactorily approximated by a uniform distribution. A common feature of many classical significance tests is that a sharp null hypothesis is compared with a diffuse alternative hypothesis. Often evidence that, for a Bayesian statistician, strikingly supports the null hypothesis leads to rejection of that hypothesis by standard classical procedures. The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing termination of data collection are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.

6. Edwards, W., "Probabilistic Information Processing by Men, Machines and Man-Machine Systems," in Proceedings of the XVIIth International Congress of Psychology (Washington, August 23, 1963), North Holland Pub. Co., Amsterdam, 1964.

This is a speech covering much the same materials as the immediately following reference; a three-page abstract of the speech will be published in the proceedings of the Congress.

7. Edwards, W., Phillips, L. D., "Man as Transducer for Probabilities in Bayesian Command and Control Systems," in G. L. Bryan and M. W. Shelly (Eds)., Human Judgments and Optimality, Wiley & Sons, New York, 1964.

This chapter, a more recent discussion of PIP, presents a fairly specific proposal for the design of a class of systems which, by using human judgment in a rather unconventional way, should be able to make more nearly optimal decisions than do present

systems intended for the same purpose. It supports the proposal by reporting an experiment which shows that men required to draw conclusions from fallible data do it poorly enough to leave room for vast improvement.

8. Edwards, W., "The Design and Evaluation of Probabilistic Information Processing Systems, Proceedings of the Fifth National Symposium on Human Factors in Electronics, May 5-6, 1964, San Diego, California, Professional Technical Group on Human Factors in Electronics, Institute of Electrical and Electronics Engineers, 1964, pp. 169-181.

A major task of a command and control system often is to determine what is happening in its environment. Conclusive information is usually lacking, so such systems must attempt to synthesize thousands of items of information, each individually worth little, into an accurate picture or diagnosis of the relevant environment. Current systems (e.g., the NORAD Combat Operations Center) use sophisticated display and information retrieval devices, but leave to unaided human judgment the task of synthesis followed by decision.

The ideas of Bayesian statistics offer the basis for a new technology of diagnostic information processing. In the Bayesian view, probabilities are orderly or consistent opinions, and Bayes's theorem of probability theory is the optimal rule for revising opinion on the basis of information. The crucial input to Bayes's theorem is the probability, for each datum to be processed and for each hypothesis of interest, that the datum would occur if the hypothesis were true. Research suggests that experts can estimate such probabilities, or numbers that can be translated into them, with fair accuracy. Once such probabilities are available, a desk calculator or computer can easily synthesize them into a posterior distribution that gives the current probability of each hypothesis of interest on the basis of all the available data.

Details of the design of such a probabilistic information processing system (PIP) are presented. Laboratory research completed and in progress is reviewed, along with simulation studies intended to compare PIPs with traditional information processing systems in complex and realistic environments.

9. Edwards, W., "Optimal Strategies for Seeking Information: Models for Statistics, Choice Reaction Times and Human Information Processing," J. Math. Psych., 1965, in press.

Models for optional stopping in statistics are also normative models for a variety of tasks in which subjects may purchase risk-reducing information before making a decision. This paper develops a Bayesian model for optional stopping in the continuous case with two hypotheses; it takes explicit account of cost of information, values of the possible outcomes of the final decision, and prior probabilities of the hypotheses. Extensive tables of numerical solutions to the model's transcendental equations are provided.

Two models for choice reaction time are derived. One is based on the normality assumptions of signal detectability theory; the other is nonparametric. They are formally identical; in this case the normality assumptions are superfluous. The nonparametric model makes strong predictions about times and errors; it has only one quantity not directly observable.

A second example uses the nonparametric model to design and predict results of a binomial information-purchase experiment.

10. Slovic, S. P., Value as a Determiner of Subjective Probability. Unpublished doctoral dissertation, University of Michigan, Ann Arbor, 1964.

The purpose of this study was to explore the manner in which judged probabilities of events are influenced by the desirability of these events.

Subjects were shown five bags, each containing 100 poker chips. They were told that one bag contained 30 red chips, one contained 40, one 50, one 60, and one 70; the remaining chips in each bag were blue. Subjects could not tell which bag was which. One of the bags was selected by the subjects and the experimenter proceeded to draw a sample of 50 chips from it, one at a time, with replacement. Subjects observed the sample and, at various times, made direct probability estimates for each of the five possible compositions of the bag. They were told that a monetary payoff would be given to them, regardless of their probability estimates, depending on what the true contents happened to be. The table below shows the assignment of payoffs to bags.

	True Composition of Bag				
	30 Red	40 Red	50 Red	60 Red	70 Red
Group I and Group I_R	\$ 0	\$ 0	\$ 0	\$ 0	\$ 0
Group II and Group II_R	lose \$1	lose \$5	\$ 0	win \$5	win \$1
Group III and Group III_R	lose \$5	lose \$1	\$ 0	win \$1	win \$5
Group II_w	lose \$1	lose \$5	\$ 0	win \$5	win \$1

Groups I, II, III, and II_w constituted Experiment I. Group II_w differed from Group II by having received a brief warning not to allow the values to bias their estimates. None of these groups were rewarded for the accuracy of their probability estimates. Groups I_R, II_R, and III_R constituted Experiment II. These groups were rewarded for accurate estimation. Groups I and I_R were control groups for whom all hypotheses had neutral desirability. A trick device enabled the experimenter to draw the same sample of chips for every group.

The results indicated that the value of an event does affect judgments about its probability. However, the nature of value biases is rather complicated. It varies systematically among subjects and among trials. Some subjects in the payoff groups were optimistic. They consistently gave higher probabilities to the desired events and lower

probabilities to the undesired events than did subjects in the control groups. Others were generally pessimistic. Despite the consistency of individual differences, value groups showed more optimism (or pessimism) at some times during the sampling than at others. These differences among trials were similar in both experiments.

The reward for accuracy did not reduce value biases. Some subjects in Groups II_R and III_R overestimated the probability of the most undesired event so that, if it did occur, the larger reward for accuracy would reduce their loss.

Bayes's theorem provides a normative model for probability estimation in this task. Probabilities given by subjects in the control groups were closer to Bayesian probabilities than were those given by subjects for whom payoffs were associated with the events. The inferiority shown by members of value groups did not diminish as they accumulated more information about the bag, and was not reduced by rewards for accuracy.

The brief warning given to Group II_W effectively reduced value biases. These subjects behaved more like those in Group I than like those in Groups II and III.

REFERENCES

1. B. R. Philip, "Generalization and Central Tendency in the Discrimination of a Series of Stimuli," Can. J. Psychol., 1947, Vol. 1, pp. 196-204.
2. S. S. Stevens and E. H. Galanter, "Ratio Scales and Category Scales for a Dozen Perceptual Continua," J. exp. Psychol., 1957, Vol. 54, pp. 377-409.
3. E. H. Shuford, "Percentage Estimation of Proportion as a Function of Element Type, Exposure Time, and Task," J. exp. Psychol., 1961, Vol. 61, pp. 430-436.
4. G. H. Robinson, "Continuous Estimation of a Time Varying-Probability," Ergonomics, 1964, Vol. 7, pp. 7-21.
5. W. H. Teichner, "Psychophysical Concepts of Probability," Psychol. Rep., 1962, Vol. 10, pp. 3-9.
6. W. Edwards, "The Theory of Decision Making," Psychol. Bull., 1954, Vol. 51, pp. 380-417.
7. W. Edwards, "Behavioral Decision Theory," Ann. Rev. Psychol., 1961, Vol. 12, pp. 473-498.
8. W. Edwards, "Subjective Probabilities Inferred from Decisions," Psychol. Rev., 1962, Vol. 69, pp. 109-135.
9. W. Edwards, H. Lindman, and L. J. Savage, "Bayesian Statistical Inference for Psychological Research," Psychol. Rev., 1963, Vol. 70, pp. 193-242.
10. W. G. Cochran and G. M. Cox, Experimental Designs, 2nd ed., Wiley, New York, 1957.
11. F. Mosteller and P. Nogee, "An Experimental Measurement of Utility," J. polit. Econ., 1951, Vol. 59, pp. 371-404.
12. M. G. Preston and P. Baratta, "An Experimental Study of the Auction-Value of an Uncertain Outcome," Am. J. Psychol., 1948, Vol. 61, pp. 183-193.
13. R. M. Griffith, "Odds Adjustments by American Horse Race Bettors," Am. J. Psychol., 1949, Vol. 62, pp. 290-294.
14. H. C. A. Dale, "A Priori Probabilities in Gambling," Nature, 1959, Vol. 183, pp. 843-845.

DISTRIBUTION LIST

Copy No.	Addressee
1-50	Electronics Systems Division Air Force Systems Command, USAF Laurence G. Hanscom Field, Bedford, Massachusetts ATTN: Contract No. AF 19(604)-7393
51-72	Scientific and Technical Information Division (ESTI) Laurence G. Hanscom Field, Bedford, Massachusetts
73	Air University AUL Maxwell Air Force Base, Alabama

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified.)

1 ORIGINATING ACTIVITY *(Corporate author)*

Engineering Psychology Lab.
Institute of Science and Technology
University of Michigan, Ann Arbor, Michigan

2a REPORT SECURITY CLASSIFICATION
Unclassified

2b GROUP n a

3 REPORT TITLE

Human Processing of Equivocal Information

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Final Report

5 AUTHOR(S) *(Last name, first name, initial)*

Edwards, Ward

6 REPORT DATE	7a TOTAL NO OF PAGES	7b NO OF REFS
April 1965	80	14

8a CONTRACT OR GRANT NO	9a ORIGINATOR'S REPORT NUMBER(S)
AF 19(604)-7393	
b PROJECT NO: 4690	3780-23-F
c	9b OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d	

10 AVAILABILITY/LIMITATION NOTICES

Qualified requestors may obtain from DDC
Copies available from OTS.

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY
None	Decision Sciences Laboratory Hq ESD, L. G. Hanscom Field, Bedford, Massachusetts

13 ABSTRACT

This report contains a series of studies investigating the abilities of subjects to revise probability estimates on the basis of new information. These studies show that subjects' probability estimates are reliable, but deviate considerably from posterior probabilities calculated from Bayes's theorem. These deviations are almost always in the conservative direction, i.e., low Bayesian probabilities are overestimated, and high ones are underestimated. Only when each datum is very ambiguous do subjects' estimates become more extreme than Bayesian probabilities. Further, when subjects are asked to give 90% or 50% credible intervals of a posterior probability distribution, their estimates are wider than Bayesian credible intervals. This finding of conservatism has led to the design of a man-computer system that should minimize the effects of human shortcomings in making diagnoses.

DD FORM 1 JAN 64 1473

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information Processing Systems						
Industrial Engineering						
Bayesian						
Decision Making						
Experimentation						
Design						
Probability						
Theorem (Bayes)						
Game Theory						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13 ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS); (S); (C); or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14 KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.